CUK-H05963-6-KP742

6

THE JOURNAL OF PHILOSOPHY

THE JOURNAL OF PHILOSOPHY was founded in 1904 as THE JOURNAL OF PHILOSOPHY, PSYCHOLOGY AND SCIENTIFIC METHODS by Frederick J. E. Woodbridge and J. McKeen Cattell. In 1906, Wendell T. Bush became associated with the JOURNAL as co-editor. The title was changed in 1921, and the JOURNAL was incorporated in the State of New York under its present name in 1923. Since its founding, it has been published from the Department of Philosophy at Columbia University and printed by Cadmus Journal Services in Lancaster, Pennsylvania.

Phin

THE JOURNAL OF PHILOSOPHY

Calcutta University Library

VOLUME XCV, NUMBER 1
JANUARY–DECEMBER 1998

Edited by
Bernard Berofsky, Akeel Bilgrami
Arthur C. Danto, Kent Greenawalt
Isaac Levi, Mary Mothersill

Calcutta University Library

Published by The Journal of Philosophy, Inc.

KP-742
(19.4.2001).

STATEMENT OF COPYRIGHT POLICY

We would like to draw your attention to the protection copyright affords against multiple photocopying by libraries, instructors, and copy centers. The 1976 copyright law provides that photocopies in excess of "fair use" copying cannot be made without authorization by the copyright owner—"fair use" meaning single copying for personal use. Accordingly, the JOURNAL authorizes personal or educational multiple copying only if a request is made in writing, and the fee of $.05 per page/per copy is paid directly to us. *Each copy must include a notice of copyright.* We shall turn over to the authors any substantial payments (over $30.00, an administrative fee) obtained from all photocopying. Inquiries about photocopying may be sent to the Managing Editor.

© Copyright 1998 by the Journal of Philosophy, Inc ISSN 0022-362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices. Send address changes to THE JOURNAL OF PHILOSOPHY, 709 Philosophy Hall, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, NY, NY 10027.

THE JOURNAL OF
PHILOSOPHY

Calcutta University Library
ALIPORE CAMPUS

VOLUME XCV, NUMBER 1
JANUARY 1998

page
5 *Conceptual Similarity across Sensory and Neural Diversity: The Fodor/Lepore Challenge Answered* Paul M. Churchland
33 *Natural Kinds and Crosscutting Categories* Muhammad Ali Khalidi
51 MEMO TO AUTHORS
52 NOTES AND NEWS

Published by The Journal of Philosophy, Inc.

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Isaac Levi, Mary Mothersill. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Philip Pettit. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

1998

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Akron and Lancaster, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1998 by the Journal of Philosophy, Inc.

ISSN 0022-362X

THE JOURNAL OF PHILOSOPHY

VOLUME XCV, NUMBER 2
FEBRUARY 1998

page
53 *Collapse of the New Wave* Ronald P. Endicott

REVIEW ESSAYS
73 Michael Sandel: *Democracy's Discontent: America in Search of a Public Policy* Philip Pettit

BOOK REVIEWS
97 Joseph Raz: *Ethics in the Public Domain: Essays in the Morality of Law and Politics* Avishai Margalit

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Isaac Levi, Mary Mothersill. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Philip Pettit. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

1998

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Akron and Lancaster, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1998 by the Journal of Philosophy, Inc. ISSN 0022-362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

THE JOURNAL OF
PHILOSOPHY

VOLUME XCV, NUMBER 3
MARCH 1998

page
105 *Biological Universals and the Nature of Fear* Mohan Matthen

COMMENTS AND CRITICISM

133 *Taking the Measure of Doom*
Alexander D. Scott and Michael Scott

BOOK REVIEWS

142 Thierry de Duve: *Kant after Duchamp* Daniel Herwitz
149 Dieter Henrich: *The Unity of Reason: Essays in Kant's Philosophy* and *The Course of Remembrance and Other Essays on Hölderlin* Fred L. Rush, Jr.

160 NEW BOOKS

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of curren interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Isaac Levi, Mary Mothersill. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Philip Pettit. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

1998

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Akron and Lancaster, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1998 by the Journal of Philosophy, Inc. ISSN 0022-362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

THE JOURNAL OF
PHILOSOPHY

VOLUME XCV, NUMBER 4
APRIL 1998

page
161 *Language Conventions Made Simple* Ruth Garrett Millikan
181 *How Great a Good Is Virtue?* Thomas Hurka

BOOK REVIEWS
204 Jarrett Leplin: *A Novel Defense of Scientific Realism*
Husain Sarkar

210 NEW BOOKS

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Isaac Levi, Mary Mothersill. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Philip Pettit. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

1998

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Akron and Lancaster, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1998 by the Journal of Philosophy, Inc. ISSN 0022-362X
POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

THE JOURNAL OF PHILOSOPHY

VOLUME XCV, NUMBER 5
MAY 1998

page
215 *Why Is a Wing Like a Spoon? A Pluralist Theory of Function*
Beth Preston

COMMENTS AND CRITICISM
255 *Multiple Meanings and the Stability of Content*
Eric Margolis and Stephen Laurence

BOOK REVIEWS
264 Lee Smolin: *The Life of the Cosmos*
Steven Weinstein and Arthur Fine

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Isaac Levi, Mary Mothersill. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Philip Pettit. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

1998

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Akron and Lancaster, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1998 by the Journal of Philosophy, Inc. ISSN 0022-362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices

THE JOURNAL OF PHILOSOPHY

VOLUME XCV, NUMBER 6
JUNE 1998

page

269 *Consensus, Stability, and Normativity in Rawls's Political Liberalism* Larry Krasnoff

293 *Property Counterparts in Ersatz Worlds* Mark Heller

BOOK REVIEWS

317 Onora O'Neill: *Towards Justice and Virtue: A Constructive Account of Practical Reasoning* Sarah Williams Holtman

322 NEW BOOKS

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Isaac Levi, Mary Mothersill. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Philip Pettit. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

1998

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Akron and Lancaster, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1998 by the Journal of Philosophy, Inc.

ISSN 0022-362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

THE JOURNAL OF PHILOSOPHY

VOLUME XCV, NUMBER 7
JULY 1998

page
323 *On Ascribing Beliefs: Content in Context* William W. Taschek
354 *Time and Mind* Andy Clark

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Isaac Levi, Mary Mothersill. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Philip Pettit. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

1998

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Akron and Lancaster, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1998 by the Journal of Philosophy, Inc.

ISSN 0022-362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

THE JOURNAL OF PHILOSOPHY

VOLUME XCV, NUMBER 8
AUGUST 1998

page
377 *Names and Expressions* Robert Fiengo and Robert May
410 *What Is So Bad about Contradictions?* Graham Priest

BOOK REVIEWS
427 Vivian Walsh: *Rationality, Allocation, and Reproduction* Daniel M. Hausman

Published by The Journal of Philosophy, Inc.

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Isaac Levi, Mary Mothersill. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Philip Pettit. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

1998

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Akron and Lancaster, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1998 by the Journal of Philosophy, Inc.

ISSN 0022-362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

THE JOURNAL OF
PHILOSOPHY

VOLUME XCV, NUMBER 9
SEPTEMBER 1998

page
THE WOODBRIDGE LECTURES 1997:
Having the World in View: Sellars, Kant, and Intentionality
John McDowell
431 Lecture I: *Sellars on Perceptual Experience*
451 Lecture II: *The Logical Form of an Intuition*
471 Lecture III: *Intentionality as a Relation*

492 MEMO TO AUTHORS

CALCUTTA UNIVERSITY LIBRARY
ALIPUR
CAMPUS

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Isaac Levi, Mary Mothersill. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Philip Pettit. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

1998

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Akron and Lancaster, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1998 by the Journal of Philosophy, Inc. ISSN 0022-362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

THE JOURNAL OF PHILOSOPHY

VOLUME XCV, NUMBER 10
OCTOBER 1998

page
493 *Proof and Truth: Through Thick and Thin* Stewart Shapiro
522 *Proper Names, Names, and Fictive Objects* Avrum Stroll

BOOK REVIEWS

535 D. M. Armstrong: *A World of States of Affairs* Alex Oliver
540 Isaac Levi: *For the Sake of the Argument: Ramsey Test Conditionals, Inductive Inference, and Nonmonotonic Reasoning* Michael Morreau

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Isaac Levi, Mary Mothersill. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Philip Pettit. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

1998

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Akron and Lancaster, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1998 by the Journal of Philosophy, Inc. ISSN 0022-362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

THE JOURNAL OF PHILOSOPHY

VOLUME XCV, NUMBER 11
NOVEMBER 1998

page
547 *The Problem in Twentieth-century Philosophy* Jerrold J. Katz

COMMENTS AND CRITICISM

576 *Preston on Exaptation: Herons, Apples, and Eggs* Daniel C. Dennett

BOOK REVIEWS

581 Robert Kane: *The Significance of Free Will* Alfred R. Mele
585 Richard Rorty: *Achieving Our Country: Leftist Thought in Twentieth-century America* David Bromwich
591 Karl Löwith: *Nietzsche's Philosophy of the Eternal Recurrence of the Same* Bernard Reginster
598 NEW BOOKS: TRANSLATIONS

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Isaac Levi, Mary Mothersill. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Philip Pettit. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

1998

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Akron and Lancaster, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1998 by the Journal of Philosophy, Inc.

ISSN 0022-362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.

THE JOURNAL OF PHILOSOPHY

VOLUME XCV, NUMBER 12
DECEMBER 1998

page
599 *Cantorian Abstraction: A Reconstruction and Defense* Kit Fine

COMMENTS AND CRITICISM
635 *On the Perfectly General Nature of Instability in Meaning Holism* Kelly Becker

BOOK REVIEWS
641 Stephen P. Stich: *Deconstructing the Mind* Patricia Kitcher
645 INDEX TO VOLUME XCV

Published by **The Journal of Philosophy, Inc.**

THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

Editors: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Isaac Levi, Mary Mothersill. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Philip Pettit. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Leigh S. Cauman; *Treasurer*, Barbara Gimbel; *Other Trustees*: Kent Greenawalt, Michael J. Mooney, Lynn Nesbit, Daniel Shapiro.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

THE JOURNAL OF PHILOSOPHY

1998

SUBSCRIPTIONS (12 issues)

Individuals	$35.00
Libraries and Institutions	$75.00
Students, retired/unemployed philosophers	$20.00
Postage outside the U.S.	$15.00

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are taken on a first come, first served basis.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Akron and Lancaster, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by consulting our January issue or by writing to us.

© Copyright 1998 by the Journal of Philosophy, Inc. ISSN 0022-362X

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices

THE JOURNAL OF PHILOSOPHY

VOLUME XCV, NO. 1, JANUARY 1998

CONCEPTUAL SIMILARITY ACROSS SENSORY AND NEURAL DIVERSITY: THE FODOR/LEPORE CHALLENGE ANSWERED

In chapter 6 of *Holism: A Shopper's Guide*, Jerry Fodor and Ernest Lepore[1] present a critical challenge to "state-space semantics," as pursued by neural network modelers. Their challenge expresses the doubt that such an approach could ever make useful sense of the notions of conceptual identity, or even conceptual similarity, in the face of the enormous functional and structural diversity across the individual networks that constitute human brains.

The present essay is the latest in a series of exchanges on this matter.[2] I stand by those earlier responses, but some new developments motivate a more ambitious and less defensive essay. Some recent theoretical insights and neuromodeling results from Aarre Laakso and Gary Cottrell[3] constitute a decisive answer to Fodor and Lepore's challenge and point to new areas of semantic research. Laakso and Cottrell successfully deploy one member of a large family of mathematical measures of conceptual similarity, measures that see past differences—even extensive differences—in the connectivity, the sensory inputs, and the neural dimensionality of the networks being compared. I shall explore the rationale behind these measures (1) by sketching Laakso and Cottrell's application of one such measure across a diversity of real networks, and (2) by discussing the general-

---

[1] Cambridge: Blackwell, 1992, pp. 187-207.

[2] My "Fodor and Lepore: State-Space Semantics and Meaning Holism," *Philosophy and Phenomenological Research*, LIII (1993): 667-72; Fodor and Lepore, "Reply to Churchland," *Philosophy and Phenomenological Research*, LIII (1993): 679-82; my "Second Reply to Fodor and Lepore," in Robert McCauley, ed., *The Churchlands and Their Critics* (Cambridge: Blackwell, 1996), pp. 278-83. But for their original chapter, the entire exchange appears in McCauley.

[3] "Qualia and Cluster Analysis: Assessing Representational Similarity between Neural Systems" (under review).

0022-362X/98/9501/5–32

ization of such measures to many-layered networks and to time-series (recurrent) networks.

I. THE PROBLEM

On neural-network models of cognition, the primary unit of representation, and the primary *vehicle* of semantic content, is the overall pattern of simultaneous activation levels across the assembled neurons of a given population, such as the cells in layer four of the primary visual cortex, or the output cells of the motor cortex, or the "cells" in a given layer of some artificial network model (figure 1). Such patterns are often referred to as *activation vectors*, because they can be usefully and uniquely characterized by a sequence or *n*-tuple of numbers, where *n* = the number of neurons in the representing population.

To those more geometrically inclined, a specific activation pattern can also be simply and usefully characterized as a specific point in a proprietary space, an *n*-dimensional space with a proprietary axis for the variable activation level of each of the *n* neurons in the representing population. Any single point in that space will represent, by way of its unique set of *n* coordinate values, the simultaneous activation levels of each of the cells in the corresponding neural population.

At this level of reckoning, a specific activation *pattern* = a specific activation *vector* = a specific *point* in activation space. These are just variant ways of talking about the same thing: the unit of representation, as physically rather than semantically described.

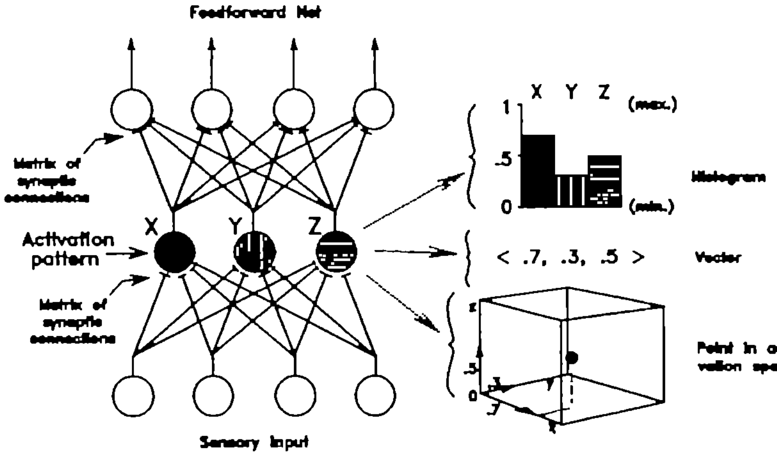


Figure 1. Three ways of conceiving of the current activity of a specific neuronal population: as a pattern or histogram of activation levels, as an activation vector, and as a point in an activation space.

If we are to assign specific semantical or representational *contents* to collective units of this kind, a natural first assumption is that any unit must in some way inherit its overall content from the individual and more basic representational significance of each of its many constituting elements, namely, the activation level of each of its many neurons. After all, it is these individual neurons which are the recipients of information from the environment, either directly—through their interaction with ambient light, heat, and various mechanical and chemical impingements, or indirectly—through their many synaptic connections with neurons earlier in the processing hierarchy.

As Fodor and Lepore have earlier urged on me, however, this assumption leads quickly to a number of difficulties. In the first place, the approach to content here emerging would seem to be just another version of a Humean concept empiricism, with the "simple" concepts corresponding to whatever content or significance is assigned to the individual axes of the global activation space, that is, to whatever content or significance is assigned to the activation levels of each individual neuron in the relevant population. A "complex" concept will then be just an appropriate conjunction or configuration of the simples, a unique configuration represented by a unique point or position within the global activation space.

What was presented as new and interesting is therefore something quite old and boring, runs the objection. Worse, the emerging account leaves unaddressed the question of how each *axis* of the activation space, or each neuron of the representing population, gets *its* semantic or representational content. Presumably, such "microcontent" must derive from the peculiar causal connections that a given neuron enjoys to microfeatures of the sensory environment, or, if the neuron in question is not itself a sensory neuron, then to some microfeatures of the content embodied in the activity of neurons earlier in the processing hierarchy, neurons that project axons to the specific neuron in question.

This obvious proposal leads in this case to disaster, however. The problem is that no two people will have exactly the same set of causal dependencies across their sensory neurons, nor, a fortiori, across their subsequent neuronal layers as we climb their respective information-processing hierarchies. Even at the initial or sensory layer, there will be considerable idiosyncracy, and that diversity will explode exponentially as the already idiosyncratic sets of microcontents get successively transformed by the profoundly idiosyncratic matrices of synaptic junctions that successively connect each layer in the neuronal hierarchy to the next layer. There is no hope, for example, of finding a mapping between the 10^{10} cells in my primary visual cortex and the 10^{10} cells in yours, such that for *every* mapped pair, your cell and mine have the *same* microcontent. The diversity in the scattered strengths or weights of our 10^{13} synaptic connections effectively precludes it. Probably, we do not share a single such pair between us.

Worse still, we are all idiosyncratic in the sheer *number* of cells we have in any given neuronal population. Our respective activation spaces, therefore, do not even have the same *dimensionality*, let alone the same set of axial microcontents. Accordingly, if we cannot make sense of the notion of "same activation space" across distinct individuals, it is plain that we can make no sense of the notion of "same position in such a space" across two individuals, or even of "proximate positions." So we have no hope of explicating the notion of "same meaning" or "similar meaning" across distinct individuals. On such an ill-behaved neural basis, to conclude, we have no hope of constructing a workable theory of meaning.

II. PAINTING A DIFFERENT PICTURE

The short answer to this critique is that content is not, in general, assigned in the manner described. A point in activation space acquires a specific semantic content not as a function of its position relative to the constituting *axes* of that space, but rather as a function of (1) its spatial position relative to all of the *other contentful points* within that space; and (2) its causal relations to stable and objective *macrofeatures of the external environment*. As we shall see, this very different focus takes our neural and synaptic diversity out of the semantic picture almost entirely. That undoubted diversity is without semantic significance. Two people can share a common hierarchy of partitions across their respective activation spaces, and, for each partition within that hierarchy, they can share a common causal relation to some objective environmental feature, despite the fact that those partitions and those causal relations are differently implemented across structurally idiosyncratic brains. Starting with a deliberately simple example, let me illustrate how this can happen.

Let us imagine that two identical feed-forward networks, #1 and #2, are trained on the same corpus of 100 photos of each of 100 members of four extended and multigenerational hillbilly families: twenty-five Hatfields, twenty-five McCoys, twenty-five Wilsons, and twenty-five Andersons (figure 2). The two networks are trained—by progressive nudgings of their respective synaptic weights at the hands of the familiar back-propagation algorithm—to distinguish any input photo as a member of one of the four families. (The input layer of each network is here graphically portrayed as a line of twelve cells, but, of course, it would have to be a two-dimensional grid of at least sixty by sixty cells, in order to accommodate a realistic picture of each face. As well, there are far too many interlayer connections to portray all of them. Forgive the graphic license here taken.) The trained network's identification of any given face is defined as the highest of the four activation levels at the four labeled output cells. The two networks, let us suppose, are eventually trained to 100% accuracy on the training set of photos, and they generalize successfully to any new examples of these four extended families with > 90% accuracy.

How do they manage to perform at this high level? In real networks, the usual answer to this sort of question appeals to the coding strategy evolved, during learning, at the middle or "hidden" layer of

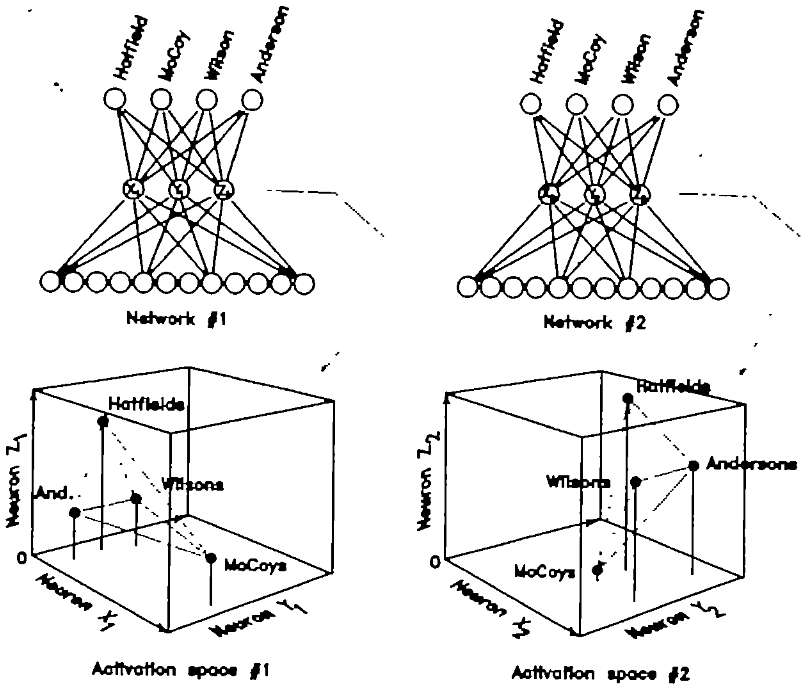


Figure 2. The locations of four prototype points within the hidden-layer activation spaces of two (imaginary) neural networks for recognizing the faces of four different extended families. The four points represent a prototypical Hatfield face, a prototypical McCoy face, a prototypical Wilson face, and a prototypical Anderson face.

neurons in the network involved. The idea is to look at the unique hidden-layer activation-space positions, severally displayed by the trained network, for grouping and separating the four families portrayed in the input photos. If the training was successful, the activation-space points for each of the twenty-five Hatfield faces will tend to be clustered together in a proprietary subvolume or partition within the overall space. The face points for other families will be similarly clustered together in distinct familial clouds, as illustrated in figure 2.

To simplify things, let us focus our attention on the mean, average, or center-of-gravity position for all of the Hatfield face points. This will be the *prototypical* Hatfield point, and it will represent the prototype of a Hatfield face—a face, perhaps, that no actual Hatfield actually possesses, but that every Hatfield resembles more or less closely. Let us plot this special point with a small bold-face sphere in the relevant space for each network, as in figure 2. Do the same for the average or prototypical face of each of the other three families. The result for network #1 is the four points illustrated to the left, and the result for network #2 is the four points illustrated on the right.

As reckoned against the axes of their respective embedding spaces, each foursome is a distinct group of points. In one obvious sense, the two networks have settled on *different* coding strategies in solution to their common discriminatory problem. But a closer look reveals a residual and robust similarity in their respective coding strategies: the *relative* positions of each familial point, as reckoned against the other three familial points in each space, are obviously very similar. Indeed, they are identical across the two networks. More generally, the linear dimensions and the vertex angles of the three-dimensional "solids" defined by those four prototypical familial points are identical in the hidden-layer activation spaces of both networks. The solid shape at issue—an irregular tetrahedron—is both rotated and translated in space #2, relative to space #1, but it is the *same* solid shape in both spaces.

What is the significance of this isomorphic configuration of points? It means that both networks have reached a common reckoning of the *relative similarities and differences* that unite and divide the 100 faces in the training set, a common reckoning of the classes (and, as we shall see, the subclasses) into which that population divides, and a common reckoning of the numerical degree of similarity that relates any two face points within the same activation space.

In the end, it is to these common groupings, and to this common framework of similarity relations across the two hidden-layer spaces, that the final or output layer of neurons in each network gradually becomes "tuned" during training. (That is why the two networks have identical, or near identical, overall performance profiles.) The bottom half of each network manages to convert the great mass of pictorial information at the input layer into the (much lower-dimensional) form of information embodied in the two hidden-layer spaces portrayed in figure 2; and the top half of the network manages to convert that intermediate form of information into the labeled pigeon holes at the output layer.

Quite obviously, the intricate matrix of synaptic connections meeting both the hidden-layer neurons and the output-layer neurons will *have* to be very different across the two networks. (If they were identical, then the hidden-layer locations and orientations of the two tetrahedrons would be the same as well, and they are not.) Equally clear, the microcontents of the three axes x_1, y_1, z_1 of network #1 must be different from the microcontents of the three axes x_2, y_2, z_2 of network #2. Those two hidden layers may each be representing the same aspects of similarity and difference across the face population at issue, but, clearly, they are distributing the "representational load" differently across the basic axes of their representational spaces. (As causally reckoned, the respective microcontents of neurons x_1 and x_2, for example, *cannot* be the same: they are connected by a *different* family of synaptic weights to identical input layers.)

What we have, then, is a pair of networks with highly idiosyncratic synaptic connections; a pair of networks with hidden-layer neurons of quite different microcontents; a pair of networks whose input-output behaviors are nevertheless identical, because they are rooted in *a common conceptual framework* embodied in the activation spaces of their respective hidden layers.

In what sense is it "the same"? In this sense: the several activation-space *distances* between the corresponding prototype positions are one and all identical, and the prototype positions identified as "corresponding" are so identified (despite their different locations in the two spaces) because each constitutes the standard causal response, at its network's hidden layer, to a typical Hatfield face (or McCoy face, and so on) presented as input to the sensory layer.

It is easy to extend this identity criterion to encompass a numerical criterion of *similarity* across such categorial frameworks. Our two tetrahedrons have six corresponding straight edges joining each of the four corresponding prototypical points. Let us ask, for each such pair of lines, AB and $A'B'$, what is the absolute value of the difference between their lengths, and then divide it by the sum of those two lengths. This will always give us a fraction between 0 and 1, where the fraction approaches 0 as the two lengths approach identity. Let us then take the average of all six of those fractions, as computed for each of the six pairs, and subtract it from 1. As expected, this measure will give us a similarity rating of 1 for the two identical tetrahedrons of figure 2. Progressively divergent shapes will get progressively lower ratings. In summary,

$$\text{Similarity} \;=\; 1 - \text{Average}\left[\frac{|AB - A'B'|}{(AB\; +\; A'B')}\right]$$

Most importantly, this simple similarity measure will be indifferent to any translations, rotations, or mirror inversions of the solid shape in question, as that shape may find itself differently located within the diverse activation spaces of diverse neural networks. What matters is not its accidental location or orientation within any given space; what matters is its internal geometry, and the causal connections of its various vertices to objective features of the sensory environment.

Still, what of the problem of comparing such categorial frameworks across activation spaces with different *dimensionalities*? Suppose, for example, that our network #2 had four hidden-layer neurons instead of only three? That very real possibility would pose no problem at all. First, a four-dimensional space can obviously *contain* a three-dimensional tetrahedral solid. Second, nothing whatever prevents its several "prototype" vertices from being causally connected to the *same* environmental features as its three-dimensional cousin. Finally, and despite expectations, the similarity measure just described can be invoked, without modification, to evaluate even these dimensionally diverse cases. Let me explain.

The important point to remember is that the basic component of this similarity measure—*distance*—has the same units of measurement, whatever the dimensionality of the space in which that distance happens to be reckoned. It matters not whether the spaces involved have different dimensions. Distances within them can always be directly compared across distinct spaces. For example, the line AB in figure 3a has the same length as line $A'B'$ in figure 3b, despite their being embedded in a two-dimensional space and a three-dimensional space, respectively; and the triangle ABC in 3c is metrically identical to the triangle $A'B'C'$ in 3d, despite the different dimensionality of the two spaces that contain them.

So long as we can measure common distances across diverse spaces, we can always specify any point configuration or hyperconfig-

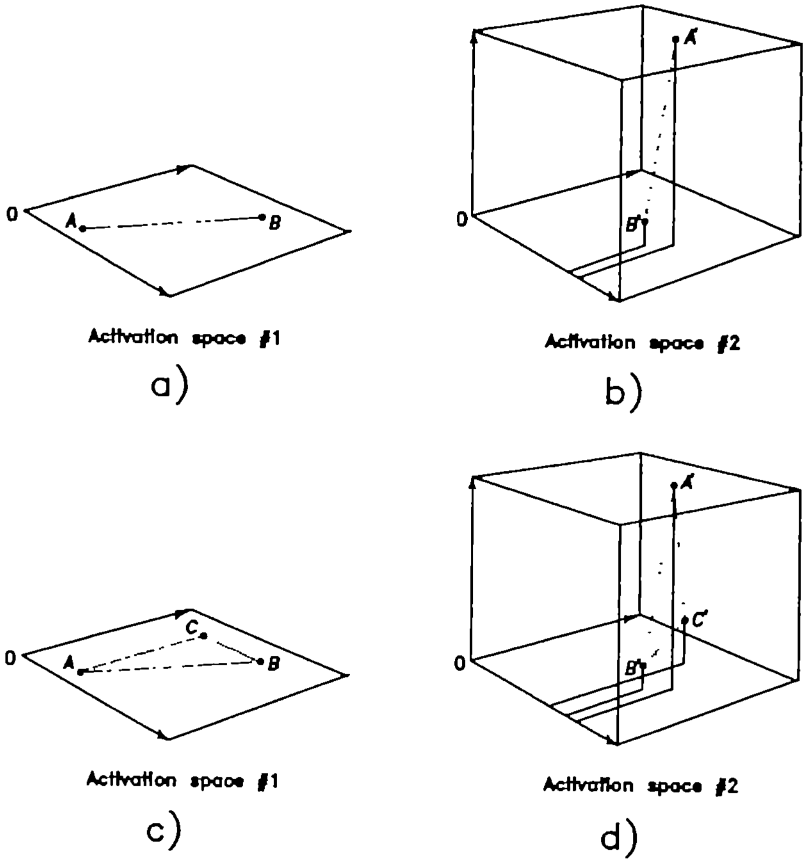


Figure 3. (a) A ten-inch line within a two-dimensional plane. (b) A ten-inch line within a three-dimensional volume. (c) A right-angle triangle in a two-dimensional plane. (d) The same right-angle triangle in a three-dimensional volume.

uration whatever, since the family of distances between its several vertices fixes that configuration uniquely (or rather, uniquely up to its mirror image). Dimensional differences, it emerges, matter not at all.

One can see this same point in another way. If a pair of dimensionally-diverse spaces both contain, in some orientation or other, a same-shape *n*-dimensional hypersolid ("same-shape" as reckoned above), then both spaces must each either constitute or contain an *n*-dimensional subspace, an *n*-dimensional hyperplane of some orientation or other, that confines each hypersolid at issue. Since those respective hyperplanes, at any rate, have the same dimensionality, there will be no problem in comparing the two solids. Simply superimpose the two (sub)spaces, as one might lift the plane of figure 3c into the space of figure 3d so as to coincide with the plane defined by its embedded triangle $A'B'C'$. If the contained shapes are indeed identical, then some relative rotation and/or translation of those two spaces will make every edge and every vertex of the two solids coincide perfectly.[4] Once again, dimensional differences simply do not matter. They seemed to, at the outset, only against the false assumptions, concerning how conceptual similarity must be reckoned, embraced by Fodor and Lepore.

III. SOME PRECURSORS OF THESE IDEAS

Laakso and Cottrell were not the first to assemble systematic distance measures in order to explicate important elements of conceptual structure. T. J. Sejnowski and C. R. Rosenberg[5] confronted a scatter of seventy-nine prototypical points in the activation space of NETtalk's hidden layer, and they decided to ask how those assembled points paired off when each was coupled to its closest neighbor. They then asked how the resulting population of pairs clustered together when closest pairs were subsequently coupled into foursomes. Repeating this process to completion resulted in a tree-like structure whose trunk divides into two main branches (the grapheme-to-vowel transforms below, and the grapheme-to-consonant transforms above), and then into a series of further divisions and subdivisions, all of which grouped appropriately similar elements together while keeping dissimilar elements apart (figure 4).

Such *dendograms*, as they are called, are highly useful representations of the sorts of hierarchically structured sets of categories that so often emerge in the course of network training. They express moderately well the sorts of distance relations (between prototype points) that we have found to be so important. Locate any two disparate points at the rightmost tips of the branching tree (for example, the two grapheme-to-phoneme transforms, $s \rightarrow z$ and $t \rightarrow T$); trace their respective branches leftward until they meet (here marked with a circle); look at the corresponding value on the horizontal coordinate axis directly below that branch junction (here, $x = 2.75$); this will be (very roughly) the activation-space distance between the two points at issue. (Strictly speaking, this gives the distance between the *average* positions of all of the points in the two branch structures

---

[4] Thanks to my student, Evan Tiffany, for pointing this out to me.

[5] "Parallel Networks That Learn to Pronounce English Text," *Complex Systems*, I (1987): 145-68.

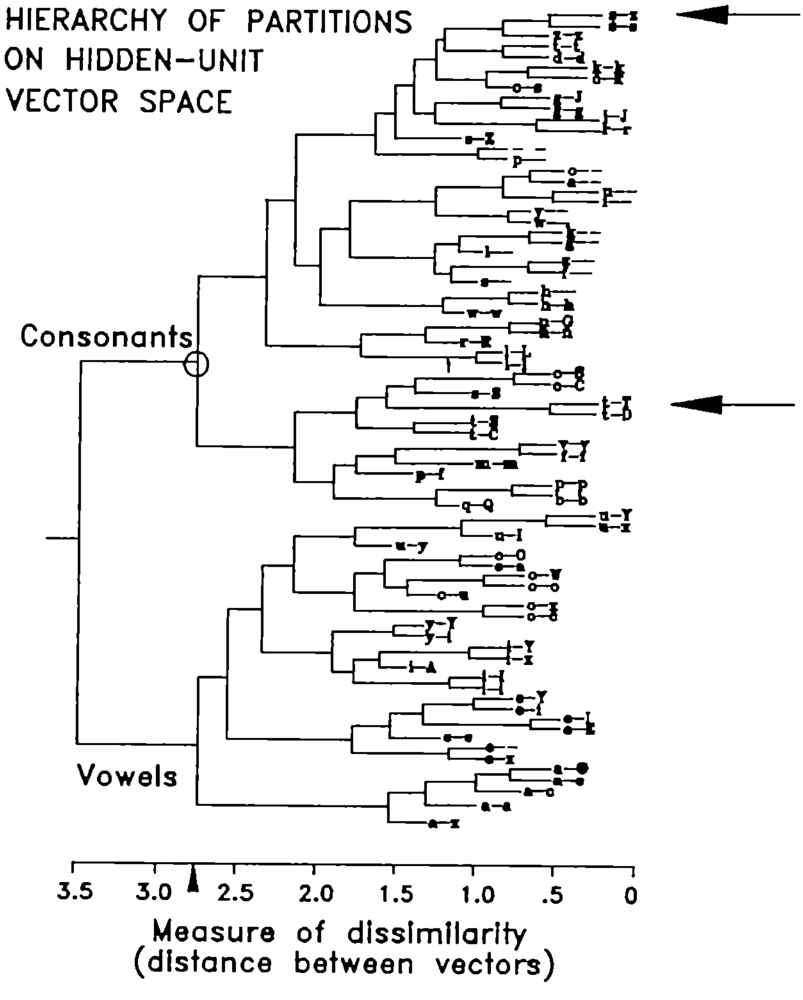


Figure 4. A dendogram representing the "closest-similarity" structure of the seventy-nine prototypical categories that emerged, during training, in the hidden-layer activation space of Sejnowski and Rosenberg's NETtalk, an artificial neural network that reads printed English text aloud.

that meet at that junction. Fine-grained distance information concerning arbitrary point pairs is thus progressively lost as one moves leftward across the diagram.)

It is worth noting that the dendogram of figure 4 retains no information concerning the *dimensionality* of NETtalk's activation space, the original space whose organization it so usefully represents. That space could have had two dimensions, or two million (in fact, it had twenty-four hidden-layer neurons). The dendogram does not care. It cares only about distances within the relevant space.

This means that dendograms are a useful device for representing both the ordinal and the metrical aspects of conceptual organization, no matter how large the representing population of neurons might be. This is not the case for literal spatial diagrams. Given the limits on our human imaginations, and on the graphic potential of a two-dimensional sheet of paper, spatial diagrams of the sort shown in figures 1-3 cease to be useful when the dimensionality of the space exceeds three. It is hard to draw an effective four-dimensional or five-dimensional figure on a two-dimensional surface. This is a grievous limitation, since many networks will have thousands or even millions of neurons at any given layer.

Nevertheless, the three-dimensional spatial diagrams are a useful metaphor for what, generally speaking, is going on the higher-dimensional cases. Specifically, there really is an abstract space and it really does come to contain prototypical points, similarity gradients, category boundaries or partitions, and a well-defined geometrical configuration that embraces all of them. The grand picture aside, however, when we wish to specify, graphically, the actual cognitive configuration of a specific network, it is to the dendogram that we must turn.

It follows from all of this that the conceptual organizations of two distinct networks, with activation spaces of quite different dimensionalities, could be correctly represented by the very same dendogram, because the relevant family of distance relations happens to be the same for both networks. This is another instance of my earlier point that there is no problem at all in specifying conceptual identity across dimensionally diverse networks.

Strict conceptual identity, however, is not what distinct empirical networks generally give us, even if they happen to be dimensionally identical, and even if they share the same set of training examples. To illustrate, let us return to the two (imaginary) face spaces of figure 2. I spoke earlier of them having a "common" reckoning of the categories involved, but a fine-grained analysis would likely show some small differences. The respective dendograms for those spaces might well look as follows (figure 5). Note that their most basic clusterings are identical, with the Wilsons and the Andersons represented as most closely similar, with both of these families represented as slightly closer to the Hatfields than to the McCoys, and with these latter two families represented as the least similar of all.

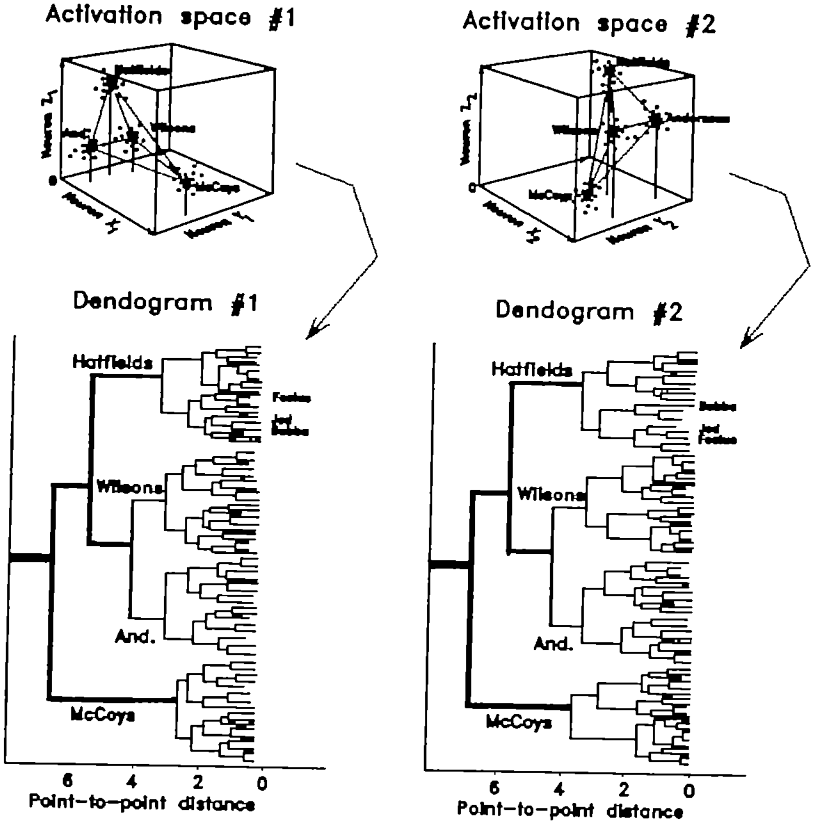


Figure 5. The respective dendograms for the two activation spaces of figure 2. Note the close similarity of their respective large-scale groupings of the 100 individual faces involved, both in who gets grouped with whom and in the relative distances that separate distinct groups. Note also the occasional small divergences in how each network pairs up "most similar" individuals at the tips of the tree's branches.

Note, however, that some of their *peripheral* pairings for most-similar faces diverge: Jed gets paired with Bubba in net #1, but he gets paired with Festus in net #2. Such residual differences in the similarity spaces of the two trained networks are typical. They reflect small differences in the initial or naive configurations of their synaptic weights, and differences in the (random) order in which they originally encountered the faces in the training set. (In human networks, they would also reflect differences in the global corpus of faces on which they were trained.) During the learning phase, the networks end up following distinct paths through the (structural!) space of possible weight configurations, and typically they arrive at quite distinct final states. Those distinct physical states may be, and presumably will be, closely similar in their functional profiles, and relevantly similar in the conceptual organizations to which they give rise, but they need not be strictly identical in either respect.

Such idiosyncracies should not be seen as a defect of our artificial models. People, too, display peripheral divergences in how they judge close similarities. What we have here is an explanatory account of how and why they might do so.

Their virtues acknowledged, dendograms have two major shortcomings. As noted earlier, they positively hide information about the exact distance between any two arbitrarily chosen prototype points (because they are progressively concerned with average positions, and with averages of averages, and so on). Across distinct networks, therefore, subtly different similarity structures can be hidden behind identical dendograms.

Second, judging the global similarity of two or more conceptual organizations may be easy for the near-identical cases, but it swiftly becomes subjective and uncertain when the internal organizations of these tree-like structures start to diverge, as often they do. As Cottrell remarked at the original presentation of the Laakso and Cottrell results: "When I eyeball the dendograms for two distinct networks, I may say, 'yes, they're fairly close', but that's just my reaction. We need an objective measure of such things."

That problem returns us to the sort of numerical similarity measure stated in the preceding section (recall figure 2), where we look at the differences between the corresponding distances that spatially configure the special family of points to be compared. That simple measure provides a similarity criterion of the objective kind required: it returns a number between 0 and 1, with 1 as a perfect score. As we saw, it is indifferent to any dimensional differences between the activation spaces being compared. But it is not the only such measure, nor perhaps the best. I have leaned on it to this point only for reasons of simplicity in exposition. Let us now look at a different but related measure, and at its application to an extended family of real empirical networks.

IV. THE LAAKSO-COTTRELL EXPERIMENTS

Laakso and Cottrell deploy a similarity measure, apparently well known among statisticians, called the *Gutman point alienation* (GPA) measure. It is applied, as in the preceding examples, to a pair of equinumerous point families, and it, too, measures the similarity in the respective internal spatial organizations of those families.[6] Furthermore, unlike my fanciful Hatfield-McCoy illustration, Laakso and Cottrell appeal to no initial clusterings around pre-averaged prototype points. They mean to apply the GPA to *all* of the actual points or activation vectors that appear in the "awake, behaving network," not just to the "prototype" points around which they may happen to cluster. This broader focus yields a much finer-grained analysis of the structure of the various similarity gradients that characterize the two spaces being compared.

Deploying this measure is perhaps the least of Laakso and Cottrell's achievements. What stands out is the series of training experiments, and the resulting family of over two dozen perceptually-skilled artificial neural networks to which the GPA was subsequently applied. It is one thing to argue, as I do in section II above, that there can be perfectly sensible measures of the kind denied by Fodor and Lepore. It is quite another to test such a measure against a large and revealing

---

[6] The GPA measure proceeds by calculating, for each of the two point families to be compared, the distance between every pair of points within that family. Those distances are then placed in rank order, from the longest to the shortest. The critical question answered by the GPA measure is: "How similar are the respective rank orderings, of the assembled distances, across the two point families?" Bluntly: "Do corresponding line segments (for example, the line segments joining the Hatfield prototype to the McCoy prototype in the two spaces of figure 2) have the same *rank* within their respective cohorts?" An identical ranking (a perfect correlation) yields a similarity measure of 1. An inverse ranking (a perfect anticorrelation) yields a measure of -1. Random correlations will distribute around a measure of 0.

The GPA has the disadvantage of deliberately losing metrical information, because it eventually turns its attention to comparing only the *rank orderings* of the respective activation-space distance families across the two networks. For this reason, I am inclined to think that a similarity measure of the sort suggested in section II will provide a more penetrating measure of conceptual similarity. Unlike the GPA, that measure does not lose metrical information in computing its assessment. This virtue comes with what may be a residual vice, however: as it stands, the metrical approach (unlike the GPA) is *not* blind to differences in *scale*. Two hypersolids, shape-identical but different in absolute size, will receive a low similarity rating. If we wish, this unwelcome sensitivity to global scale is easily repaired. Simply sum all the lengths of the first hypersolid, and divide it by the sum of all of the lengths of the second hypersolid, to obtain a "correction factor," c, thus:

$$c = \Sigma\,(AB)\ /\ \Sigma\,(A'B')$$

Then insert that correction factor into the original metrical measure, as follows:

$$\text{Similarity} = 1 - \text{Average}\left[\frac{|AB - c(A'B')|}{(AB + c(A'B'))}\right]$$

This measure will give a similarity rating of 1 for two hypersolids whose *relative* internal distances are all identical, but which differ globally in absolute scale.

family of empirical examples. That is what Laakso and Cottrell have done.

The target problem addressed by Laakso and Cottrell was not facial recognition, but color recognition. They began with a data base for 627 distinct colors, a data base that provided a unique twelve-element spectral-reflectance vector for each one of those 627 colors. Those reflectance vectors were suitable for entry, as activation patterns, across the "sensory" neurons of the model network in figure 6. It has twelve input units, three hidden units, and five output units, whose function is to code for each of the five broad color *categories*—red, yellow, green, blue, and purple—to which the network was slowly trained. It learned to group correctly, into the categories imposed, the many elements in the test population of 627 examples.

In fact, Laakso and Cottrell successfully trained *five* individual networks, of this same configuration, to a criterion level of competence on the color-recognition task. Each network differed only in the initial random setting of its synaptic weights, and in the random order in which the training examples were presented to it. (As usual, this yields a distinct learning excursion, in structural space, for each network.) After competence had been reached, no two networks shared the same final configuration of synaptic weights, despite the identity of their acquired discriminatory competence. The question then was, what had happened to the internal organization of the hidden-layer activation space in each network? In particular, had those activation spaces settled on the *same* organization, in the "family-of-corresponding-distances" sense that is by now becoming familiar?

The GPA was set to work comparing the geometrical similarity of the relevant point families (families with 627 members each, instead of just four, as in the Hatfield-McCoy example), and its answers were strongly positive. Across this set of five structurally similar networks, the lowest similarity rating returned for an activation-space pair was .93. The average pairwise similarity across this group was .97.

(A qualification needs to be made here. Computing the pairwise distances for every point-pair in a set of 627 points is computationally demanding: there are (627 x 266)/2, or 196,251 such distances. For a single pair of networks you have to do it twice—once for each activation space—and then you have to compare all of the results. As well, it may be overkill, since the GPA measure is ultimately interested only in the rank ordering of the relevant distances. Choosing discretion over valor, and finesse over brute force, Laakso and Cottrell settled for a partial computation of the all-up GPA. The point-to-point distances actually computed and compared for rank ordering were a subset of the total: those lines which shared a common vertex. This partial-GPA measure is thus less thorough than its parent, but it is still a very stern measure. Laakso and Cottrell employed it for every example in what follows, so convey the qualification forward.)

The emergence of such close similarities is intriguing. But Laakso and Cottrell went several steps further. In addition to the familiar *twelve*-element decimal-fraction coding vectors described earlier, they concocted a second, and a third, and a fourth scheme for encoding the same information contained in the original 627 reflectance vectors. The second coding scheme used a binary rather than a decimal code, and it yielded input vectors with *ninety-six* elements. The third coding scheme used five gaussian filters (rather like the human retinal system, but with five "cones" instead of three) and it yielded input vectors with *five* elements. The fourth used a simple sequential encoding from 001 to 627, and it yielded input vectors with *three* elements each.

Each sensory coding scheme, of course, required a structurally distinct network with a numerically distinct input layer in order to accommodate the four different forms in which the color information was coded. As they had done for the twelve-element case, Laakso and Cottrell constructed a family of five identical networks with a ninety-six-unit input layer, five further networks with a five-unit input layer, and five more with a three-unit input layer (figure 6). These further networks correspond to creatures with importantly *different sense organs* for apprehending and coding color information. They give us a total of twenty networks in all.

The various networks in these further families were subjected to the same training regime as the five networks already discussed. These new networks also reached high levels of performance (though the nets using the crude three-element scheme were somewhat weaker than the others). As well, they each settled into an idiosyncratic configuration of synaptic weights. Once again, the question was: What of the internal organizations of their hidden-layer activation spaces?

The GPA returned high values in measure of their mutual similarity. Within networks using the ninety-six-element scheme, the lowest similarity rating was .89 and the average similarity was .95. For the five-element networks, the values were: lowest, .94, and average, .98. For the three-element networks, the values were: lowest, .99, and average, .99. These results parallel the outcome of the first experiment. They show, once again, that similar nets facing similar problems have a robust tendency to settle into the same abstract so-

KP-742

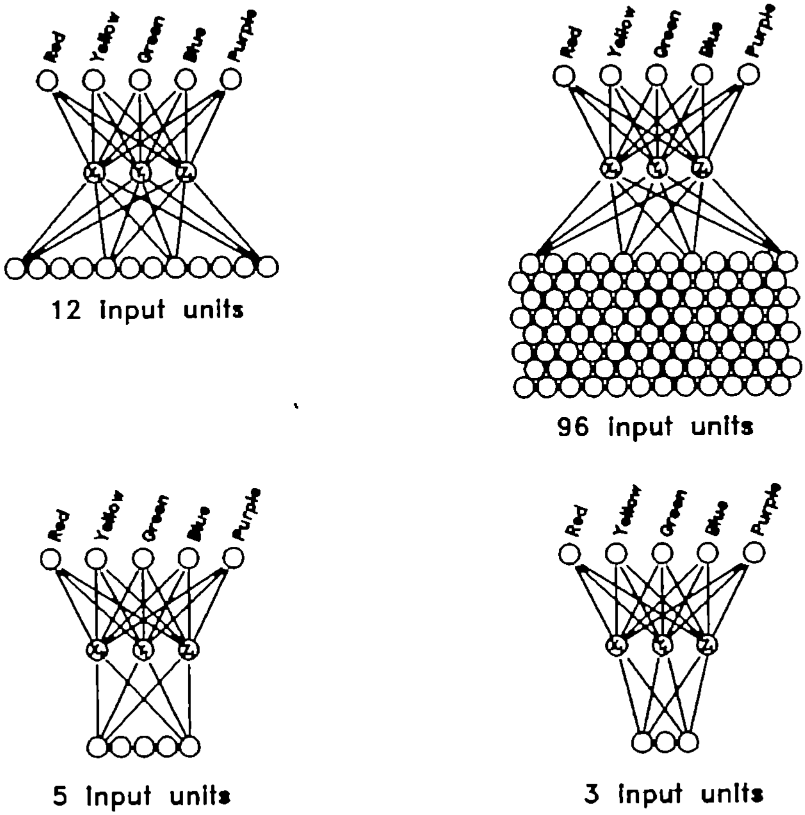


Figure 6. The architectures of the four main network types trained to color-recognition competence by Laakso and Cottrell. Note the differences in their respective sensory or input layers, and the corresponding differences in the number of synaptic connections that meet their respective hidden layers.

lution in how they structure the partitions or prototype families within their hidden-layer activation spaces.

Digging deeper, however, Laakso and Cottrell discovered something even more interesting. They set the GPA to measuring the mutual similarity of the internal organizations of the hidden-layer activation spaces, not within, but now *across* these four (structurally diverse) families of networks. Of the ten pair-wise comparisons between the ninety-six-element and the twelve-element networks, the average similarity was .95. This is the same high level of similarity that they found *within* each family of networks. The substantial structural diversity across these two families, and the diverse character of their respective sensory inputs, apparently made no difference to the activation-space solution into which they severally settled.

Other pairings across the other network families yielded comparably high measures of mutual similarity. Only when the "deliberately marginal" three-element scheme was made party to the comparisons did the measure occasionally fall as low as .72, and even here it averaged about .8. These particular results are striking, because they show that close conceptual similarity—as we are here measuring it—can and does emerge from networks with a different structural organization, and even from networks with distinct *sense organs*. (This last result is especially welcome to some of us. Readers of *Scientific Realism and the Plasticity of Mind*[7] will recall that the possibility of driving the very same conceptual framework with radically different sense organs was one of the major themes of the second chapter.)

In summary, so long as the relevant information is somehow implicit in whatever sensory-input schemes happen to be employed, and so long as the training procedures impose the same requirements on recognitional performance, then diverse networks can and often will settle into almost identical abstract organizations of the similarity gradients across their hidden-layer activation spaces. During learning, that is, the networks of the four different types come to "see past" the structural idiosyncracies of their diverse sensory manifolds, so as to achieve a common grasp of the objective features of the world that drive their several but similar behaviors. Once again,[8] and give or take a needless heaven, this is not "Hello, David Hume"; it is "Hello, Plato."

In a final series of networks, Laakso and Cottrell explored the consequences of expanding, beyond three, the *number* of hidden-layer neurons in such networks. The point here was to address Fodor and Lepore's final worry about networks whose hidden layers define spaces of different dimensionalities. Accordingly, a final family of networks, with variously-expanded hidden layers, were trained to competence on the same discrimination task posed to the earlier networks, and the GPA was used to measure the similarity in the conceptual organizations that emerged during training in their several activation spaces (figure 7). Once more the GPA returned numbers in the range of .9 and higher, the dimensional diversity notwithstanding. The networks all had different synaptic configurations—meaning that the relevant point families must, at the least, have

---

[7] New York: Cambridge, 1979.

[8] Compare my "Second Reply to Fodor and Lepore," p. 282.

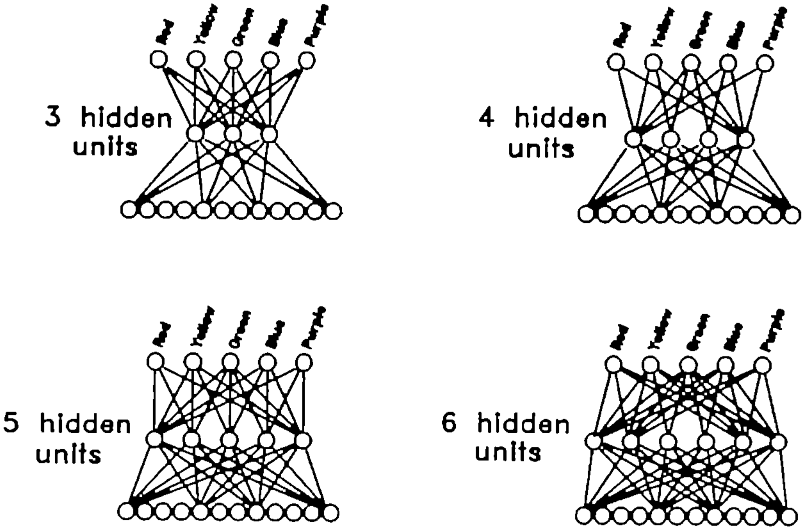


Figure 7. The architectures of four further color-recognition networks that have progressively larger numbers of hidden units. Despite the dimensional diversity of these four distinct representational spaces, the GPA discovered, after the networks were trained, nearly-identical conceptual organizations within each.

different orientations in the several spaces at issue—but here, too, the networks were found to have settled on the same conceptual organization in response to their common discriminatory problem.

The take-home point of the preceding pages is not that the various networks always settled on the *same* cognitive configuration in response to their shared problem. These networks did, and that fact is interesting. But networks need not and do not always show such uniformity in the "conceptual" solutions they find to common learning problems. The truly important point is that we can *tell* whether or not they did. We can say *what* their internal cognitive similarity consists in, and we can give an objective numerical *measure* of that similarity. Thanks to Laakso and Cottrell, these are no longer purely a priori points. We now have over two dozen trained networks of whom to ask such questions, and, taking these examples pairwise, we have at least (24 x 23)/2 = 276 distinct empirical applications of the similarity measure at issue. The results overturn Fodor and Lepore's claim that a useful measure of this kind is impossible.

V. FUTURE DIRECTIONS

These developments are encouraging, but we should remember that this is still an early skirmish in what will surely be a long campaign.

The full range of semantic phenomena includes far more than the hierarchy of partitions we have discovered within the hidden-layer activation spaces of three-layer, feed-forward, classifier networks. Our discussion, so far, stops there. Although I can do little more than scout the territory, I shall close this essay by looking ahead to what remains to be done.

Multi-layer feed-forward networks demand attention first. Few creatures have but a single neuronal layer between their sensory and their motor neurons. How, then, should we reckon conceptual similarity across networks whose input and output layers are separated not by one hidden layer, but by many? Here, I think, it is relatively easy to generalize the story of the preceding sections.

Let us assume, as before, that we are comparing two classifier networks whose learned input-output functions are close to being identical. Their gross discriminative behaviors, that is, are the same. In that case, the most obvious place to look for a pair of similarly-partitioned activation spaces is at the penultimate layer of each network, since the output layer of each must base its discriminative verdicts on exactly the information that appears in the immediately preceding layer. If we want to know "where it all comes together," therefore, the *last* of the networks' hidden layers is the presumptive place to look for each network's acquired conception of how the objective world is structured.

This presumption is encouraged by the lesson of the four different network architectures trained to the same competence by Laakso and Cottrell. The diversity across those networks lay in their earliest layers, a diversity that had disappeared by the time we reach their last (and only) hidden layers. The point is generalizable. Two networks can have many layers of hidden units preceding their penultimate layers, and yet come to exactly the same cognitive configurations within their penultimate activation spaces. Those earlier layers can be as idiosyncratic as you please, and the number of such layers may well differ from one network to another. But this need not impede the relevant application of the GPA or other similarity measure to the learned point families within each of the two penultimate spaces. The results of such reckoning will have the same significance they had in the original three-layer cases.

We can, of course, apply the GPA so as to compare the internal structures, across the two networks, of *any* of their corresponding layers. Simply generate a corresponding family of points in each, by presenting the "same input" to each network (as we saw, it may be differently coded), and turn the GPA loose. It may discover a series of close similarities up and down the two hierarchies in question, or it may not. (Laakso and Cottrell actually did this for the respective input layers of their four distinct networks: the similarities here were dramatically lower.) What to make of the results of such comparisons will presumably vary from case to case, but we can be assured that we can always address such questions and hope to get a determinate answer. Multilayer networks, to conclude, seem not to present any insuperable problems, either in conceiving of similarity or in measuring it.

Recurrent networks, by contrast, launch us into a new universe of subtleties. The first major difference is structural: recurrent networks are characterized by the existence of at least some axonal projections that reach back from their originating neuronal population to make synaptic connections with some hidden layer that is earlier or below it in the processing hierarchy (figure 8). The result is a hidden layer whose activity is driven partly by the upward flow of sensory information, and partly by the downward flow of some already-processed information, information that is a digest of some of the network's immediately preceding activity.

Functionally, this arrangement allows for an ongoing modulation of the network's cognitive response to any sensory input, a modula-

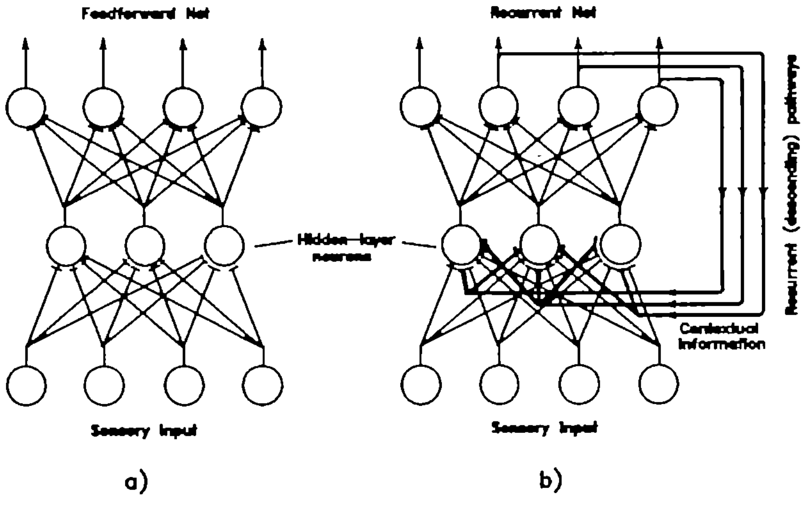


Figure 8. (a) A purely feed-forward network. (b) A recurrent network. Note the flow of information from layer three back to the neurons in layer two. This allows the network to modulate its own responses to sensory inputs, and to profit from thus recalling its own immediate history.

tion driven by a form of short-term memory of the network's proximate cognitive past. Such modulation can continue to drive the network even if its sensory layer should happen to fall silent, for the recurrent information can be quite sufficient to keep the ball rolling, once it is started.

I have elsewhere discussed at length the cognitive properties of recurrent networks,[9] and shall not attempt to retell those stories here. What is important for present purposes is that a recurrent network's primary unit of representation is not the *point* in activation space, but rather the *trajectory* in activation space. It is a temporally extended *sequence* of the now-familiar activation patterns across the hidden layer. The virtue of such trajectories is that they can represent objective phenomena with a *temporal* profile, such as a running animal's gait, a bouncing ball's path, or an unfolding sentence's grammatical structure. Feed-forward networks represent the world as an unrelated sequence of timeless snapshots or frozen spatial structures. But recurrent networks represent the world as unfolding in time, as something that contains prototypical processes, structured sequences, and standard causal pathways.

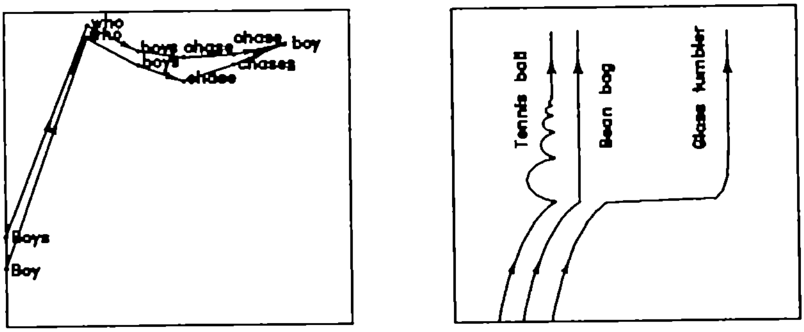
Let me first address the similarity of such representations as it is reckoned *within* a given network's hidden-layer activation space. As expected, and as figure 9a illustrates, similarity consists in the spatial proximity of the trajectories being compared. The two possible trajectories there pictured represent two possible sentences that differ only in the number of their subject and in their final verb: "Boys, who boys chase, chase boy," and "Boy, who boys chase, chases boy." Jeffrey Elman's[10] well-known grammar network yielded identical trajectories in its representation of grammatically identical sentences entered as input. It yielded proximate trajectories in its representation of grammatically similar sentences, and it yielded distant trajectories when representing grammatically diverse sentences.

Note that similarity can be restricted to temporal segments of the compared trajectories. Figure 9b illustrates three trajectories that represent, respectively, a falling tennis ball that bounces several times, a falling bean bag that flattens on first impact, and a falling glass that shatters on impact. Each of the three objects is represented

---

[9] *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science* (Cambridge: MIT, 1989), chapters 10, 11. See also my "Second Reply to Fodor and Lepore."

[10] "Grammatical Structure and Distributed Representations," in S. Davis, ed., *Connectionism: Theory and Practice*, in *Vancouver Studies in Cognitive Science*, Volume III (New York: Oxford, 1992), pp. 138-94.



Sentences

a)

Projectiles

b)

Figure 9. (a) The activation-space trajectories of two grammatically similar sentences in Elman's recurrent "grammar" network (*op. cit.*). (b) A family of trajectories that are similar until a critical time, at which point they diverge sharply.

by a spatially distinct activation-space trajectory, because the three objects are quite rightly comprehended as objects with distinct causal properties. Despite the proximity of their initial (falling) phases, therefore, they have divergent post-impact paths. Another example of this sort of restricted similarity would be the two trajectories for two sentences whose first six words were identical, but whose grammatical form diverged from the seventh word on.

We are here beginning to refine our similarity judgments to address *partial* similarities, or similarity in certain specifiable aspects. This bids me point out a second and broader way in which this can be done. Since specific hyperplanes (within the same overall activation space) regularly code for specific aspects of the global phenomenon being represented, we can make principled judgments of partial similarities between two trajectories by looking at their partial images within a specific subspace. Two trajectories may be closely proximate in one subspace, and yet be quite distant when reckoned in another. (A rough analogy here is comparing the various partial derivatives of a pair of high-dimensional functions: their slopes may be identical in one plane, but different in another.)

Elman's grammar network once more provides an illustration of this point. Similarity and difference of relative-clause structures were most saliently represented in one tilted hyperplane, and similarity and difference in noun and corresponding verb number (singular or plural) were represented in another. All of this allows us to say that two trajectories can be quite sensibly compared and contrasted in a number of distinct dimensions.

But what of the primary problem here? What of comparing trajectories across distinct networks? Here, as with the purely feed-forward case, we must be mindful of the enormous neuronal and synaptic diversity in the ways in which the "same" conceptual framework can be implemented in two distinct individuals. Here, as there, what determines the semantic content or representational significance of a given trajectory will be the set of activation-space relations it bears to all of the other contentful trajectories within its home space, and the causal relations it bears to objective processes in its external environment.

The most obvious criterion of cross-person conceptual similarity for recurrent networks, accordingly, is a direct temporal generalization of the criterion we used in the feed-forward case. Two networks have the same conceptual organization if and only if there is some rotation, translation, and/or mirror inversion of the prototype-trajectory family of the first network such that, when the space (or relevant subspace) of the first network is projected onto the space (or relevant subspace) of the second, all of the corresponding trajectories (as identified by what sensory inputs activate them) coincide perfectly.

I am moderately confident that this is a sufficient condition for conceptual identity, but I have a small reservation concerning its credentials as a necessary condition. The main worry is one I have discussed before, though in a different context.[11] Evidently it is possible for two people (for the sake of illustration, suppose they are Isaac Newton and Christian Huygens) to have identical conceptions of ballistic-particle phenomena, and identical conceptions of travelling-wave phenomena, and yet differ substantially in the sensory phenomena that serve spontaneously to activate those conceptions in each of them. When Newton turns his attention to the phenomenon of light, it is his ballistic-particle prototype trajectories that standardly get activated in his brain. But when Huygens turns his attention to the same range of phenomena, it is his travelling-wave prototype trajectories that standardly get activated. Plainly said, Newton is convinced that light is a stream of particles and Huygens is convinced that light is a wave train.

This case poses a problem for the criterion just outlined, because the presumed identity of conceptual machinery across these two individuals *fails* the requirement that identical concepts must have the same causal connections to the sensory environment. Here, they do not. The sets of connections overlap, to be sure—Newton and Huygens agree on billiard balls, comets, water waves, and sound—but they also show resolute divergences when our two heroes confront light.

---

[11] *A Neurocomputational Perspective*, chapter 11.

The lesson of this example, and we can contrive many more, is that we should weaken the requirement that demands sameness of sensory causal connections. At the very least, we should demand such sameness across individuals only for the majority of environmental domains, and allow diverse responses within others.

For these sorts of reasons, I have always been moved to play down the semantic importance of concept-to-world connections, and moved to play up the semantic importance of the set of all concept-to-concept connections.[11] It is the sameness in the latter, across two individuals, that most decisively marks them as having the same conception of the world in which they live. This impulse toward a "purely-internal conceptual-role" theory of meaning was present in my *Scientific Realism and the Plasticity of Mind*, and it moves me still, despite the substantial reconception of the basic kinematics of cognitive activity that connectionist models have introduced. I must own, however, that the considerations that motivate this tilt—namely, the desire to account for misconceptions of the phenomena, broad theoretical disagreements, and systematically false perceptual judgments—come prominently into play only when we are considering the cognitive activity of highly sophisticated intellectual creatures. For simpler creatures, it may be mostly idle to worry that the two constituting threads of semantic identity—concept-to-concept relations, and concept-to-world relations—will ever pull in opposing directions. Let me put this worry aside, then, at least for the time being, and return to a "two-aspect" approach to semantic identity, at least as a first approximation to the truth.

I make this conciliatory move for a further reason. A neglected element in the reckoning of semantic content is the set of causal connections that join cognitive states to *motor behavior*. If we have motivation for a two-aspect semantic theory, we have equal motivation for a *three*-aspect theory. It shows up even within folk-psychological reckonings of content, for, errors aside, we judge the content of an executed "intention" by the motor behavior it causes. This third aspect is highly salient from our new activation-space perspective, since in real creatures, perhaps the primary *function* of activational trajectories is to generate a corresponding sequence of motor behaviors. The administration of motor behavior is almost certainly the original function of activational trajectories, evolutionarily speaking. Neural ganglia that direct feeding behaviors, excretory behaviors, heartbeats, and basic locomotion probably antedate perceptual ganglia that track environmental events. They are in any case of equal importance to the life of any individual. We do not usually think of our internal cognitive activities as representing complex bodily behaviors—it is the perceptual end of the system that tends to dominate our philosophical attention—but most assuredly they do represent such things, and any proper account of cognitive representation in general must include the motor end of the system as well.

---

[11] See, for example, the short critiques of W. V. Quine, Donald Davidson, and Hilary Putnam in my *Scientific Realism and the Plasticity of Mind*, section 9; and the longer critique of Fodor in my "Perceptual Plasticity and Theoretical Neutrality: A Reply to Jerry Fodor," *Philosophy of Science*, LV (1988): 167-87; reprinted in *A Neurocomputational Perspective*, pp. 255-79.

I close by addressing a skeptical worry from the preconnectionist semantic tradition. Specifically, what makes the preceding story of activation patterns, structured spaces, and coincident trajectories a *semantic* theory? Why is it not just a new and unorthodox account of our underlying "syntactic" activities?

The answer, insofar as I can give one, goes something like this. The account we are currently piecing together is indeed a *syntactic* account, but it is not *just* a syntactic account; for it promises to do what we have always expected a semantic theory to do. Specifically, it gives us an account of how distinct representational contents are systematically embodied in distinct cognitive vehicles. It gives us a detailed account of how those contents are causally related to proprietary aspects of the external world, both in perception and in action. It shows us how certain aspects of the world's structure are mirrored in the relational structure of our assembled representations; that is, it gives us some idea of how our background conceptual framework and specific activations within it *picture* both the global structure of the world and specific things and events within it. Furthermore, the theory gives us a detailed account of how those representations participate in a principled internal cognitive economy. It also provides a criterion for assigning the *same* contents to the representational vehicles of distinct individuals. It gives us, that is, a criterion for accurate translation across the representational/cognitive systems of distinct individuals. It may well be false, but it looks like a semantic theory to me.

It also looks like something else: a theory of perceptual *qualia*. In fact, and as their title and their classification task suggest, Laakso and Cottrell first launched their investigation out of a concern for qualia, not out of a concern for concepts or universals. As well, I have long been puzzled, and a little troubled, that connectionist models fail to mark any obvious difference-of-kind between qualia and concepts. Some time ago, it occurred to me that perhaps the models were trying to tell us something: that there *is no* sharp or principled distinction between them. What we are pleased to call a perceptual quale is simply the activation of a conceptual category embodied in a hidden layer that is maximally close to the sensory transducer end of the processing hierarchy. What we are usually pleased to call our abstract concepts (including our concepts *of* such qualia, should we happen to have developed any) are just categories embodied in some hidden layer rather higher up in the same hierarchy. If so, the connectionist approach can lay claim to an unexpected unification in our understanding of cognitive phenomena.

PAUL M. CHURCHLAND

University of California/San Diego

NATURAL KINDS AND CROSSCUTTING CATEGORIES*

There are many questions that one can ask about categories in science and in common sense, and there are many ways of construing the claim that some categories are more "natural" than others. One can ask whether a system of categories is innate (for example, up/down) or acquired by learning (bourgeoisie/proletariat), whether it is theoretically based (vertebrate/nonvertebrate) or ad hoc (under one kilogram/over one kilogram), whether it pertains to a natural phenomenon (plant/animal) or to a social institution (legal/illegal), whether it is lexicalized in natural language (red/blue) or requires a compound linguistic expression (red-and-round/blue-and-square), and whether it is projectible (green/blue) or nonprojectible (grue/bleen). These distinctions are all relevant to the question of whether a certain system of categories is *natural* or *artificial*. The existence of such divergent construals of what makes a category or system of categories *natural* renders suspect any univocal answer to this question in any particular case.

Yet another question one can ask which some authors take to have a bearing on the issue of the *naturalness* of categories is whether a system of categories constitutes a unique way of organizing a particular set of entities or phenomena, or whether there are other legitimate classification schemes that can coexist with it. Another way of putting this is by asking whether systems of categories can cut across one another, and if so, under what circumstances. Some philosophers have claimed that crosscutting systems of categories cannot exist as genuine natural kinds. Richmond Thomason[1] has claimed that natural kinds are arranged in a hierarchy, such that higher categories in the taxonomic system do not trespass on the boundaries among the categories at the lower levels. In other words, two kinds can only overlap if one of those kinds is wholly subsumed under the other. So an individual can belong to two or more kinds, provided they can all be put in subsumption relations with each other. The

---

* For comments and criticism, I am especially grateful to Tom Nickles, John Heil, and members of a 1996 National Endowment for the Humanities Summer Seminar on the metaphysics of mind. Earlier versions also benefited from the comments of Murat Aydede, Michelle Mason, Jesse Prinz, Terry Regier, Richard Rosenblatt, Jessica Wilson, and audiences at the American University/Beirut and the University of Maryland/Baltimore County.

[1] "Species, Determinates, and Natural Kinds," *Noûs*, III (1969): 95-101. For some criticisms, see my "Carving Nature at the Joints," *Philosophy of Science*, LX (1993): 100-13.

0022-362X/98/9501/33–50

natural-kind categories should form a nested hierarchy of categories which are *disjoint*, or, in my terminology, which do not *crosscut*.

Peter Gärdenfors[2] is another contemporary philosopher who has advocated the idea that natural kinds do not cut across one another. As part of an attempt to respond to Nelson Goodman's "new riddle of induction," he has argued that natural properties are those which do not crosscut our existing scientific conceptual scheme. Using the standard Linnaean taxonomy in biology as his illustration, he points out that the biological classification scheme can be represented by a tree, in which a property is natural only if it applies to all parts of the classificatory tree which lie below one particular node in the tree. Relative to this tree, all other properties are nonnatural. "For example," he writes, "the properties 'marsupial' and 'vertebrate' will be natural properties in the phylogenetic classification, while 'featherless' and 'biped' will not" (*ibid.*, p. 127). This position leads him to rule out grue-like properties as nonnatural, thereby enabling him to propose a solution to Goodman's riddle.

Similarly, Ian Hacking[3] has articulated the view that scientific categories are arranged in taxonomic hierarchies. While acknowledging that some scientific categories cut across others, he has posited that these crosscutting categories are not "real kinds" (a notion borrowed from John Stuart Mill). Hacking uses this line of argument to explicate Thomas Kuhn's idea of incommensurable scientific theories. On his interpretation, the reason that some systems of categories are incommensurable with others is that they cut across each other. Since crosscutting is not allowed among scientific categories that are "real kinds," classification schemes that do crosscut are to be considered incommensurable rivals.

---

[2] "Induction, Conceptual Spaces, and AI," in Douglas Stalker, ed., *Grue: The New Riddle of Induction* (Chicago: Open Court, 1994), pp. 117-34. Gärdenfors's position raises a question as to the original provenance of our existing scientific conceptual scheme, which allegedly cannot be crosscut. His view is that it has an evolutionary origin. My concern here is not with whether Gärdenfors's move helps to solve Nelson Goodman's riddle, but rather with whether it is a reasonable way to distinguish natural from nonnatural scientific categories.

[3] "Working in a New World: The Taxonomic Solution," in Paul Horwich, ed., *World Changes: Thomas Kuhn and the Nature of Science* (Cambridge: MIT, 1993), pp. 275-310. Hacking cautions that some of Thomas Kuhn's writings upon which he is relying are marked: "Draft: not for distribution, quotation or paraphrase," and suggests that these may not be Kuhn's considered opinions (p. 307n). In what follows, I shall not be directly concerned with whether Hacking's is a reasonable construal of Kuhn's views, or of incommensurability, but with whether theories with crosscutting categories should generally be regarded as rivals, or merely as ways of organizing the same set of phenomena that are capable of coexisting. It is not clear whether Hacking endorses these views wholeheartedly or whether he expounds them merely on Kuhn's behalf. Be that as it may, it is not important that these views should be either Kuhn's or Hacking's, for they have considerable appeal for philosophers who are concerned to make the distinction between natural and nonnatural categories or kinds. Other philosophers have regarded this thesis as implicated in the notion of natural kinds. For a supporter, see Douglas Browning, "Believing in Natural Kinds," *South West Journal of Philosophy*, IX (1978): 135-48; Browning writes: "Things may be sorted into numerous overlapping types, but not into several natural kinds, unless they stand in a hierarchical arrangement, as of species to genus to higher genus" (p. 136). For a dissenter, see Ronald de Sousa, "The Natural Shiftiness of Natural Kinds," *Canadian Journal of Philosophy*, XIV (1984): 561-80.

In addition to its role in explicating the notion of a natural kind, there are two connected reasons to examine the claim that bona-fide categories do not crosscut. The first is that this claim is tied to the further claim that there is one best classification scheme in each area of human inquiry. As I have already suggested, this assumption of uniqueness is clearly related to the idea that there cannot be alternative classification schemes that trespass upon one another. The second reason for focusing on this thesis is that the injunction against crosscutting is involved in one influential derivation of essentialism. The assumption that natural categories are hierarchical seems to be needed in arguing that each such category has a unique essence associated with it across possible worlds.[4] For all these reasons, the thesis that natural categories do not crosscut is in need of further elucidation and examination.

In what follows, I shall often refer to the position that I am opposing—that natural categories cannot crosscut one another—as the *hierarchic thesis*. Before proceeding, it is worth distinguishing a strong and a weak version of the hierarchic thesis. There are those who might argue, more weakly, that categories cannot crosscut only within a single theory, as opposed to across the entirety of science or our entire collection of theories. The authors mentioned do not seem to have this weaker claim in mind. For one thing, both Thomason and Hacking envisage (apparently independently) carving up the world into the categories: *animal*, *vegetable*, and *mineral*. This system of categories, which is supposed to pave the way for noncrosscutting subdivisions, does not pertain to a single theory. Having said that, there appears to be nothing objectionable about the weaker claim that some self-contained systems of categories fail to be crosscutting.[5] The Linnaean system in biology is the most notable such system of nonoverlapping categories. Biologists and philosophers of biology commonly label such systems *hierarchies*; they are also sometimes referred to as *taxonomic systems*. 'Hierarchy' is the term used by George Gaylord Simpson[6] to refer to nonoverlapping classificatory schemes, and he reserves the term 'key' for a classification system that contains within itself crosscutting categories, though he adds that this is not the usual understanding of the word. I shall argue here that whereas hierarchies or taxonomic systems may exist within particular scientific theories, crosscutting abounds among the categories of science and lack of crosscutting is not a good criterion for the naturalness of a category.

---

[4] I argue elsewhere (*op. cit.*) that the hierarchic thesis is implicated in an influential argument for essentialism. This suggests that the claim that natural kinds pick out essential properties is ultimately dependent on the idea that natural kinds are hierarchic.

[5] Although the weaker claim is prima facie plausible, there is a possible difficulty with it. Someone who makes the weak claim that categories cannot crosscut *within* a given theory can attempt to deal with purported counterexamples simply by pronouncing them to pertain to a different theory. If this move is not to be ad hoc, however, it must rely on an independently motivated account of the demarcation of theories, an account that is not readily available.

I. CROSSCUTTING

Hacking has put forward a definition of a scientific *taxonomy*, which is his term for a hierarchy of nonoverlapping, nonsubdividing categories that culminate in a set of basic categories. He states that a taxonomy is determined by a class of entities and a transitive asymmetric relation. Such a pair, consisting of a class *C* and a relation *K*, is a taxonomy if and only if: first, *C* has a head, a member of the class that does not stand in relation *K* to any other member, but such that every other member stands in this relation to it; and second, every member except the head stands in the relation *K* to *some* other member. Hacking also says that such taxonomies sometimes terminate in a bedrock of basic categories that cannot be further subdivided (*op. cit.*, pp. 286-89). To make this more concrete, suppose that *C* is the class of biological taxa, and *K* is the relation of lower taxa to higher taxa—say, species to genus, or family to order. Then, *C* has two heads, the animal kingdom and the plant kingdom, neither of which stands in relation *K* to any other taxon. Moreover, every other taxon in the phylogenetic scheme is related to some higher taxon as species is to genus. The taxonomic claim is tantamount to the idea that scientific categories are arranged in a hierarchy, such that higher categories in the taxonomic system do not trespass on the boundaries among the categories at the lower levels. In Hacking's terminology, scientific categories cannot overlap; in my terminology, scientific categories cannot crosscut. I prefer to use the terminology of crosscutting instead of overlapping, since superordinate categories that wholly include others might be said to overlap with their subordinate categories but not to crosscut them.

Hacking conjectures that scientific categories belong to taxonomic hierarchies in order to explicate the Kuhnian claim of incommensurability.[7] As reported by Hacking, Kuhn wants to use the hierarchic thesis to explain why incommensurability arises. According to Hacking, whenever we have two scientific taxonomies, their categories either overlap one another, or the categories of one taxonomy subdivide the lowest categories of the other, or else their categories coincide. In the first two cases, overlapping and subdividing, Hacking uses the hierarchic thesis to rule out the possibility of translation. In the case of overlapping categories, there cannot be a translation between new and old, since they belong to different taxonomies. In the case of subdivision, the kind in the old science is a category with no scientific subkinds, "so the old name cannot be translated into any expression in the new science that denotes a scientific kind" (*op. cit.*, p. 295). Only in the third case, when categories coincide, can there be translatability.

---

[6] *Principles of Animal Taxonomy* (New York: Columbia, 1961), p. 13.

Thus, Hacking offers a concrete reason as to why two rival scientific theories may not be translatable one into another or both into a common language by appealing to the notion of hierarchic scientific taxonomies. To support this claim further, he cites Fred Sommers[8] as having proposed a theory of predicates in natural language which is equivalent to this one. Sommers's theory of types holds that our ontological system is arranged in a hierarchy of nested types. Among the least inclusive types is the category of persons, which are the only things that can be *angry*, or can be *philosophers*. More inclusive is the category of physical things, which consists of the types of things that can be *red* or *heavy*. The most inclusive type is the category of things that can be *thought about*, or can be *interesting*, since these predicates can be applied to virtually anything in our ontology. Note that the entities included in each type or category are not the ones to which these predicates actually apply, but ones to which the predicates are applicable, which is to say that it would not be a category mistake to apply them.

---

[7] Although he uses it here to support Kuhn's views on incommensurability, Hacking's attitude toward the hierarchic thesis is more ambivalent in other work. In one article, he notes that "most of what we readily recognize as natural kinds do have this [taxonomic] feature," and goes on to ask whether this feature is "just part of our ancient predilection for natural history," or whether it is, more significantly, "somehow a fact about the functionally relevant groupings in nature"—see "Natural Kinds," in B. Barrett and R. Gibson, eds., *Perspectives on Quine* (Cambridge: Blackwell, 1990), pp. 129-41, here p. 133. Elsewhere, however, Hacking casts doubt on the related idea of a unique taxonomy of nature: "The idea of a complex exhaustive taxonomic framework does not make sense, not even as an ideal to which we strive"—see "A Tradition of Natural Kinds," *Philosophical Studies*, LXI (1991): 109-26, here p. 111.

[8] "Types and Ontology," *Philosophical Review*, LXXII (1963): 327-63.

Sommers's theory says that, if C_1 and C_2 are any two categories, then either C_1 and C_2 have no members in common or C_1 is included in C_2 or C_2 is included in C_1. That is, if a predicate is applicable to some entities of a certain type, and some of another type, then one of those types must be wholly contained within the other. No predicates are applicable to some entities of one type and some but not all of another type. Now, it might seem that 'tall' is such a predicate. Some things that can be true can be tall (for example, stories) but not all (for example, theories). At the same time, some things that can be tall cannot be true (for example, buildings). But this just shows that the predicate 'tall' is ambiguous when applied to stories and buildings. When the ambiguity is eliminated, 'tall' can be interpreted as two different predicates applicable to different types of entity. Eliminating the ambiguity eliminates the crosscutting.

Sommers's scheme appears to be corroborated by some empirical evidence presented by cognitive psychologists concerning language learning. If so, that would make the theory of hierarchic categories something more than a claim about the predilections of scientists and their ways of classifying things; it would make it a fundamental cognitive constraint on human classification schemes generally. That is how the psychologist Frank Keil[9] conceives it in his research on language learning in children. Keil thinks that the principle of not having crosscutting predicates is obeyed throughout the language-learning process by young children. Even though such children sometimes misapply predicates, they almost never consider predicates applicable to some members of a certain type and not to others of that same type. He postulates that only predicates that obey this principle are "psychologically natural." As in Sommers's scheme, whenever the constraint appears to be violated, there is some ambiguity in the predicates involved. Keil cites a variety of experiments to show that there is no crosscutting throughout psychological development, and these lead him to conclude that the principle of hierarchic types is "a constraint on human cognition" (*ibid.*, p. 165).

---

[9] *Semantic and Conceptual Development: An Ontological Perspective* (Cambridge: Harvard, 1979). The Sommers-Keil claim has been regarded by some cognitive psychologists as a basic constraint on language learning, but it has also met with criticism in those quarters. Susan Carey has come up with purported counterexamples to Keil's claim, but I shall not assess her conclusions, since for my purposes it is enough that the (stronger) hierarchic thesis is false, as I shall argue in the following section—see Carey, "Constraints on Semantic Development," in W. Demopoulos and A. Marras, eds., *Language Learning and Concept Acquisition: Foundational Issues* (Norwood, NJ: Ablex, 1986), pp. 154-71.

II. REAL KINDS

Sommers's theory of types does not strengthen the hierarchic thesis, and Keil's evidence from developmental psychology does not either. There is a crucial difference between the hierarchic account of taxonomy and the Sommers-Keil theory of types: the latter does not consider sets of things to which a predicate applies or fails to apply, but those to which it would or would not be *applicable*. A predicate is said to be applicable of something if and only if it would not be a category mistake to apply it, not if it would be merely false. Thus, 'tall' is applicable to both buildings and dwarves, though it is presumably only *true* of some buildings and no dwarves. Even if there is no crosscutting when it comes to applicability, there still may be when it comes to actual application. Sommers's claim is notably *weaker* than the hierarchic thesis because it relies on the notion of a category mistake rather than mere falsity.

While Sommers's constraint rules out the existence of two predicates that are applic*able* to crosscutting sets of things, the hierarchic thesis rules out the existence of two such predicates that truly apply to crosscutting sets. There is some plausibility to the claim that the counterexamples to crosscutting types are ambiguous (such as the predicate 'tall'), but it is easy to find nonambiguous counterexamples to the hierarchic thesis among the predicates of natural language. Consider the predicates 'red' and 'round': there is some overlap between the two sets, since some balls are both round and red, but there are also members of each set that do not belong to the other. Some nonround things are red (for example, fire engines) and some nonred things are round (for example, snowballs). This does not constitute a counterexample to Sommers's and Keil's constraint, since the Sommers-Keil theory would say that the predicates 'red' and 'round' are applicable to all three things, though they are not all correctly applied to all of them. The categories of natural language may display a hierarchical structure when it comes to applicability, but it is clear that they do not when it comes to application. Whatever the merits of the claim that the hierarchy of applicability constitutes a basic cognitive constraint, the hierarchic thesis does not have the same status when applied to the predicates of natural language. Still, it might be claimed that it holds for scientific predicates. So the question arises: Is it possible to save the claim by restricting it to genuine scientific predicates or categories?

The hierarchic thesis implies that an individual can belong to more than one scientific category, provided all these categories can be put in subsumption relations with each other. The claim is false, however, if taken to apply to scientific categories in general. To cite just one example, the category *parasite* crosscuts the category *insect*. Tapeworms and fleas are classified together as parasites, and fleas and flies are both classified as insects, but neither category, *parasite* and *insect*, includes all three. Hacking seems to admit that there can be scientific categories that crosscut one another without belonging to incommensurable theories. As an example, he gives the category *poison*, which "overlaps" the categories *vegetable* and *mineral* but is surely not incommensurable with them.[19] Although he recognizes that *poison* is a legitimate scientific category (after all, the scientific field of toxicology is based upon it), he argues that it is not a "real kind." He borrows the notion of a real kind from Mill in order to come up with a modified version of his original claim: while some scientific categories can be crosscutting, real kinds cannot. To qualify as a real kind, there must be an inexhaustible number of things to find out about a category. Whereas this is true of, say, *arsenic* in Hacking's opinion, it is not true of *poison*, since "There is nothing much common to poisons except what puts them in the class in the first place, namely the potential for killing people after being ingested" (*op. cit.*, p. 300). By contrast, he quotes Mill as saying that when it comes to real kinds, we discover "new properties which were by no means implied in those we previously knew" (*op. cit.*, p. 301). Hacking thinks this distinction between real and nonreal kinds has some application in science and defends it on independent grounds. But his reason for introducing it is that it enables him to explain why some categories from different theories are incommensurable: they are real-kind categories that belong to different taxonomic hierarchies. To put it differently, incommensurable theories are those which carve the world into real kinds that crosscut one another.

Does the notion of a real kind rescue the hierarchic thesis? Hacking admits that he does not have a proof that all real kinds belong to taxonomic hierarchies or that all taxonomic hierarchies contain real kinds. Nevertheless, he thinks that the notion of a real kind is a useful one and that hierarchic taxonomies "still carry some cachet" in the sciences (*op. cit.*, p. 303). But it is neither clear that the notion of a real kind is a useful one in science, nor that all exceptions to taxonomic hierarchies are nonreal. Crosscutting categories abound in science and they cannot all be dismissed as nonreal kinds. A notable one is the example mentioned earlier, the category *parasite*, which crosscuts the Linnaean tree in biological classification. It is clearly a real kind, because there are certain things that have been discovered to be common to all parasites that were not built into the category in the first place. To quote some typical scientific findings from a standard textbook on parasitology: "Appropriate triggering mechanisms initiate the change from infective stages to parasitic stages. Once the parasite has begun its existence in a new host body, other triggering mechanisms initiate each change of the parasite during its development."[11] Such information was not part of what was initially put into the category *parasite*. In fact, there are few scientific categories that have nothing more discovered of them than what was put in, and Hacking is right to be concerned that his claim will bring on accusations that he is committed to the analytic-synthetic distinction. A category that had such a feature would be tied to an irrevocable definition of the kind that does not survive in the context of inquiry. Moreover, it is quite clear that this cannot be regarded as a proper case of incommensurability, since these two taxonomies (associated with zoology and parasitology) coexist comfortably, whereas incommensurable theories are supposed to be rivals.

---

[19] It is unlikely that *vegetable* and *mineral* are genuine scientific categories, but I shall grant this for the sake of argument. If not, *poison* can be seen to crosscut the categories *organic* and *inorganic*.

III. INTERESTS

Let us retrace our steps. The hierarchic thesis has been proposed by several authors as a characteristic of natural kinds or natural scientific categories. These writers have thought of natural categories as being arranged in a noncrosscutting hierarchy. Hacking has conjectured that a category that crosscuts such a hierarchic taxonomy is incommensurable with that system of categories. In response to purported counterexamples, such as the category *poison*, which crosscuts certain chemical categories without being incommensurable with them, he modifies the claim so that it applies only to real-kind categories (in Mill's sense). The revised claim is that real-kind categories that overlap with taxonomic systems of real kinds are incommensurable with those systems. The revised claim is also subject to counterexamples, however: the category *parasite* clearly conforms to the characterization of a real kind. The same could be said for the entomological categories—*larva*, *pupa*, and *imago*—that cut across the Linnaean species categories, or of the phase categories—*solid*, *liquid*, *gas*—that cut across the categories of the periodic table. Or, for that matter, the chemical categories—*acid* and *base*—that crosscut the categories *organic* and *inorganic*. All of these categories are real kinds in the sense that we discovered things about them which were by no means implied when they were first introduced. Far from being incommensurable, crosscutting schemes are capable of coexisting in our total theory of the world. To say that they are *capable* of coexisting is not to say that they always *do*. Some crosscutting categories may be rejected for the same reasons that scientific categories are generally rejected—but how and why this happens is a topic for another discussion. It is quite correct to say that crosscutting taxonomies cannot be translated into one another, but incorrect to think that they are incommensurable rivals. We should not expect crosscutting categories to be intertranslatable, any more than we should expect genetics to be translatable into cosmology.

---

[11] E. R. Noble and G. A. Noble, *Parasitology: The Biology of Animal Parasites* (Philadelphia: Lea and Febiger, 1982, 5th ed.), p. 7.

But, it may be protested, there is a difference in the two cases: the relationship of genetics to cosmology is not the same as that of parasitology to zoology. There is still something of a problem in determining whether two theories are in competition with one another and can be translated into one another, or whether they merely coexist comfortably, though they share a subject matter. One needs to explain *why* there is no temptation in the case of parasitology and zoology to say that the two theories are incommensurable rivals, despite the fact that they are about some of the same entities. These two theories both concern living organisms, whereas genetics and cosmology do not. Why, therefore, are they not rivals? Their ability to coexist may be explained by saying that they embody different *interests* relating to a common subject matter. In the case of the Linnaean taxonomy, the governing interest is determining the diachronic-phylogenetic relations among organisms, while in the case of parasitology, the main interest is in determining synchronic-ecological relations among a subset of those organisms.

The idea that there are crosscutting taxonomies is closely related to the view that scientific classification is *interest relative*. If classification is always relative to certain interests, we would expect some taxonomies to reorganize some of the same entities in different ways without displacing existing ones. It is not entirely clear how to specify the exact role of interests in grounding crosscutting classification schemes, but this is what makes some of these schemes capable of cohabitation, unlike say classical mechanics and relativistic mechanics, which had roughly the *same* interests and were hence rivals. Therefore, one need not rest with the brute fact of failure of translation to explain why certain theories do not come into conflict when they have the same subject matter, whereas others do. This fact can be explained in terms of the interest relativity of scientific theories. Although it jars with some of his other claims, Hacking may also have something like this interest relativity in mind when he writes (in a parody of Karl Marx): "But, of course, I can act out different lives, work in different worlds, splice genes in the morning and go to the homeopath in the afternoon" (*op. cit.*, p. 298).

It might be objected that some *rival* theories do have different interests; indeed, that some theories are rivals precisely because they have different interests, and that one is sometimes replaced by another when the governing interests change. For example, it might be said that astronomy displaced astrology when we were no longer interested in using the motions of the stars and planets to predict human behavior and became more interested in these motions for their own sake. This is a misleading account of the relationship between astronomy and astrology, however, since these interests were intertwined for many centuries, during which time the two disciplines were effectively inseparable, until it became clear that the two sets of interests could be detached and the two disciplines eventually diverged. In fact, interest relativity may help to account for the persistence of astrology and to explain the fact that it has not been entirely supplanted by astronomy to this day.

Disparity of interests is also a feature that often distinguishes folk classifications from those of the experts. The relationship between common sense and scientific taxonomy is a topic that has elicited considerable discussion recently among philosophers, psychologists, anthropologists, and others. In the biological realm, it has been extensively explored by Scott Atran,[12] who argues that many folk taxonomies survive and exist comfortably alongside scientific taxonomies. This fact is often obscured, because we assume that the mere appearance of a scientific classification of a particular realm of inquiry leads automatically to the displacement of the common-sense classification. Atran argues persuasively that this is not always true historically, not even in the modern (supposedly scientific) era, since, even today, common-sense concepts in biology are not always identical with scientific concepts. Scientific categories and common-sense categories often coexist amicably without clashing. Thus, the transition from natural history to biology "involved not so much a radical rupture with common sense, as maintaining a continuing access through its reevaluation" (*ibid.*, p. 13). He also states: "If laypeople accept modification of a folk taxon, it is because the scientific taxon proves compatible with everyday common-sense realism; if not, the scientific concept can usually be set aside, and the lay notion persists as a 'natural kind' regardless" (*ibid.*, pp. 6-7). To this I would add that when scientific categories displace those of common sense, that is because the two subscribe to roughly the *same* interests. This proposal is borne out by cases in which common-sense classifications drop out and are replaced by the scientific ones.

---

[12] *Cognitive Foundations of Natural History* (New York: Cambridge, 1990).

Replacement of the folk taxonomy by the scientific one is more prevalent, for example, in medicine and the classification of diseases than in other areas. As Atran notes, most naive concepts of disease give way to sophisticated ones, rather than continue to exist alongside them. It may not be immediately obvious why diseases should be less resilient to scientific advances than other folk categories and why common-sense disease categories generally give way to scientific ones. Atran suggests briefly that concern with taxonomic nosologies (classification of diseases) is a specialized affair "by and large restricted to doctors and naturalists of the seventeenth and eighteenth centuries, and to twentieth-century ethnolinguistics and ethnomedicine. Most folk have no need or use for it" (*ibid.*, pp. 311-12). The fact that the treatment of disease in many cultures is left largely to experts may help to explain the relative ease with which many traditional folk diseases have dropped out and given way to scientific categories. But this does not really answer the question as much as postpone it. We can go on to ask: Why is the classification of diseases left up to the experts? Interest relativity can help answer this question: in this case, the purposes and interests of the experts are by and large the same as the folk, namely, the treatment of diseases, the alleviation of their symptoms, and the health of persons afflicted with them. Since the folk and the experts have identical or at least highly similar interests, we should not expect the folk taxonomies to survive in this case and that is why in some instances they are displaced by expert taxonomies, whereas in other cases they are coopted by the experts and further refined.

These proposals may also be related to some of the debates among philosophers of mind concerning the status of folk psychology and its eventual fate in the face of future developments in psychology. Some have argued that the folk theory will be eliminated by a scientific psychology, whereas others have insisted that it is not a rival to a science of psychology and will continue to flourish alongside it. Although that is not usually how it is framed, the central question may be put in terms of interests: Do these theories have the same or similar interests, or are they pursuing different interests? Some writers who have tackled the question have touched on this issue. Daniel Dennett's[13] distinction between three different *stances* that one can take toward the mind-brain can be viewed in this light (*ibid.*, pp. 16-17). What Dennett calls the physical, design, and intentional stances are, at least in part, ways of specifying different interests when it comes to the study of the mind-brain. From the physical stance, one predicts the behavior of a system by determining its physical constitution and the physical impingements upon it, then by using the laws of physics to predict the outcome for any input. From the design stance, the organism is treated as a device or artifact that behaves as it is designed to behave under different circumstances. From the intentional stance, by contrast, the organism is regarded as an agent that has formed rational beliefs about the environment and reasons about the world in conformity with those beliefs. Similarly, John Dupré[14] has argued that neurobiology, psychology, and folk psychology have different concerns and are interested in explaining different facets of human beings. He writes: "(Human) neurobiology, I believe, is a science concerned largely with how people work rather than with what they do; psychology has some important concerns with both kinds of question; and folk psychology is mainly concerned with understanding what people actually do" (*ibid.*, p. 152).

Therefore, crosscutting categories are not only rife among bona-fide scientific classification schemes, they are also in evidence in comparing scientific taxonomies with folk taxonomies. In both cases, the reason that categories from different taxonomies can cut across one another without being in competition is that they pertain to different interests. In the concluding section, I shall attempt to use this notion of interest to introduce the notion of a scientific domain.

IV. DOMAINS

Some philosophers, scientists, and others have thought of crosscutting systems of categories as being mutually exclusive or as pertaining to rival classification schemes. I have argued that scientific categories routinely cut across one another without being rivals. In examining these issues, we have been led to an idea that deserves more attention than it has hitherto received in the philosophy of science. That idea concerns the role of interests in delimiting the area of inquiry for scientific theories. Scientific disciplines and subdisciplines may be distinguished not only by their subject matter, but also by the *interests* relative to which they conduct their inquiries. This idea has some relevance to current discussions of objectivity, which sometimes assume that the mere presence of interests jeopardizes the objectivity of science, and use the expression 'interest theory' to describe the social constructivist view of science. The above considerations make clear that interest-free science is not an option, since all scientific theories involve a set of interests relative to which the inquiry is conducted. In other words, no inquiry is literally dis*interest*ed. Rather, objectivity is partly determined by the extent to which one lives up to the interests that one sets out to serve. But it also involves ruling on the appropriateness of those interests in the first place; for example, we might question the interest of profit making if it is the main guiding factor in pharmacological research. These considerations do not resolve our worries about objectivity; there are still difficult questions to be addressed concerning the manner of determining the interests operable in any given enterprise, the individuation of those interests, and the judgment that any particular interest or set of interests is illegitimate in a certain area of research. But this way of looking at things may be helpful in tackling the problem of objectivity.

---

[13] *The Intentional Stance* (Cambridge: MIT, 1987).

[14] *The Disorder of Things: Metaphysical Foundations of the Disunity of Science* (Cambridge: Harvard, 1993).

Most real examples from science will not be capable of neat resolution, and the interests involved will not be easily classifiable as appropriate or inappropriate. To take a thornier example, consider Donna Haraway's[15] claim that the science of animal groups or animal sociology has been "unusually important in the construction of oppressive theories of the body political" (*ibid.*, p. 11). She argues that around the time of World War II, scientists studying the behavior and social organization of primates saw their subjects in terms of dominance and other oppressive social relations. This attitude toward their subject matter derived from the experiences of these scientists and their interests, which Haraway claims were sexist and geared to promoting "scientific management of every phase of society" (*ibid.*, p. 13). Given this background, she argues that it was natural for these scientists to create categories (such as *dominance*) that tended to serve their interests. How might such biases be corrected? Haraway notes that these sciences have already been revised to some extent: for example, by switching the focus from using primates as models of human beings to a deeper look at the primates themselves. Moreover, she proposes that such sciences should be built on "social relations not based on domination" (*ibid.*, p. 19). This progression in the development of primatology can be understood in terms of a changing background of interests that evolved in the years since the World War II *away* from domination and social control and *toward* liberation and a more intrinsic interest in the primates for their own sake. Haraway advocates the alternative approach, while admitting that the alternative that she favors is also an "avowedly interested approach to science" (*ibid.*, p. 23). There is clearly more work to be done in showing that one set of interests is more appropriate than the other if we are to make further progress on the notion of objectivity in science. But casting the issue in terms of a changing set of interests helps us to avoid some of the quagmires often associated with discussions of objectivity and subjectivity.

---

[15] *Simians, Cyborgs, and Women: The Reinvention of Nature* (New York: Routledge, 1991).

It may seem as if the acceptance of the interest relativity of scientific theories will lead to the following problem. Suppose that we use it to salvage discredited theories by saying of such theories that they are not false but that they satisfy different interests. We may even specify these interests in such a way that makes the whole notion of interest relativity look suspect. For instance, a pharmacological theory might be defended by saying that it serves the interest of securing funds from a pharmaceutical company. In response, it should be pointed out that a mere appeal to interests cannot automatically save a discredited theory. Interest relativity is involved in picking out a system of categories, but those categories will still need to account for the phenomena and withstand the usual evidential, explanatory, and predictive tests. If another system of categories does the job better than the first, then we cannot simply alter our interests to save the first system. Among other things, we expect the categories that we develop to be nonvacuous, have clear criteria of application, feature in regularities, and figure in explanations. If a system of categories fails badly in all or some of these ways, it cannot be rescued simply by invoking interests. In fact, once a set of categories does not satisfy our usual standards, that may legitimately cause us to doubt the appropriateness of the interests with which the categories were picked out.

One implication of this discussion is that it is not enough to point to a set of entities or phenomena in order to specify the focus or area of inquiry of a particular discipline or subdiscipline. The domain of a theory is not picked out merely by adverting to a set of entities in some theory-neutral way, but also by the specification of certain interests relative to which one undertakes the inquiry. In fact, it is not even clear that entities can always be picked out in a theory-neutral way, since our interests are often instrumental in specifying what the entities or objects in a given domain will be.[16] From the microbiologist's point of view, there are no such entities as genes, since the chromosome is not divided neatly into discrete sequences, yet the geneticist finds it necessary to divide it thus. But since interests, stances, or perspectives seem insufficient on their own to pick out scientific domains, we need to specify an additional parameter. William Wimsatt[17] has appealed to the notion of "levels," by which he means "compositional levels—hierarchical divisions of stuff (paradigmatically but not necessarily material stuff) organized by part-whole relations, in which wholes at one level function as parts at the next (and at all higher) levels" (*ibid.*, pp. 5-6). For example, he states that neurons are composed of parts, such as membranes, dendrites, and synapses, which in turn are made of molecules, which are made of atoms, and so on. Levels "are constituted by families of entities usually of comparable size and dynamical properties, which characteristically interact primarily with one another..." (*ibid.*, p. 7). The temporal factor is also important, since processes at higher levels tend to take place at lower rates than processes at lower levels. This raises the possibility of being able to specify entities by further adverting to size and tempo, since the entities would appear to drop out once we indicate the spatiotemporal dimensions. As Wimsatt points out: theories come in levels "because that's where the entities are" (*ibid.*, p. 11).[18] In addition to levels, Wimsatt also appeals to a notion of "perspective," which is closely related to what I have been referring to as a set of interests: "Anatomy, physiology, and genetics are different perspectives on an organism..." (*ibid.*, p. 19).[19]

---

[16] A similar point is made in John Haugeland, "Pattern and Being," in B. Dahlbom, ed., *Dennett and His Critics* (Cambridge: Blackwell, 1993), pp. 53-69. In commenting on Dennett's notion of a stance (mentioned in the previous section), he writes: "But a stance is more than just an attitude toward or a perspective on things, more even than a method and terminology for dealing with them. Adopting a stance is *taking a stand*. Why? Because it is this alone—*commitment* to constitutive standards—that allows that toward which the stand is taken to stand out as phenomena, to stand over against us as objects. Such standards determine the *being* of the objects..." (p. 65). Similarly, I have argued elsewhere that two stances toward the mind issue in two different concepts of *concept*, ones that pick out different entities—see my "Two Concepts of Concept," *Mind and Language*, x (1995): 402-22.

[17] "The Ontology of Complex Systems," manuscript (University of Chicago, 1994).

[18] Wimsatt is reluctant to identify levels in a brute spatiotemporal fashion. Although he thinks that size is important to the determination of levels, he makes clear that size is not everything: "consider bacterium-sized black holes, which definitely would not exhibit Brownian motion" (*ibid.*, p. 10) But the difference in level here may be explained by a difference in temporal rates of interaction and development, since these are very different for black holes and bacteria.

Since interests are not sufficient to specify an area of inquiry, I propose to characterize a scientific *domain* as a *level* coupled with a set of *interests* (or a perspective, or a stance—for these purposes, I am treating these three notions roughly interchangeably).[20] For example, pharmacology and toxicology investigate some of the same chemical compounds and biochemical processes at the same spatiotemporal level, but whereas the interest of the first is the use of chemicals to make humans healthier, the interest of the second is to determine the properties of chemicals that have a tendency to harm humans. It is not immediately clear how to relate this mode of differentiating scientific inquiries to the more traditional ones. At first, it might seem that different disciplines investigate different spatiotemporal levels, whereas different subdisciplines investigate the same levels relative to different interests. But there are many exceptions to this rule. Some disciplines cut a swath through a wide range of levels, as in the case of physics. Within some subdisciplines, different interests may generate more than one classification scheme. In toxicology itself, poisons may be classified in terms of the target organ they affect, in terms of the chemical mechanism they exploit, in terms of their poisoning potential, in terms of their route of absorption into the body, and so on.

---

[19] Tough metaphysical issues crop up here concerning ontology and causality. I shall not attempt to address them in any detail, but they should at least be mentioned. *Ontology*: Should we construe scientific domains as ontological realms (domains of being) or merely explanatory ones (domains of description)? *Causality*: Do causal relations obtain in only one basic domain or do they obtain in all domains concurrently (the former threatens epiphenomenalism and the latter causal overdetermination)?

[20] Wimsatt would not endorse this characterization: "[L]evels come out as a kind of special case of perspectives on this analysis—a class of perspectives which map compositionally to one another so that their entities are related without cross-cutting overlaps in a hierarchical manner" (*op. cit.*, p. 19). But he says that there are also perspectives that are not levels: "They may fail to be levels because they are too small—they are located mostly at levels, but aren't of sufficiently broad span to count as levels. Or they may fail to be levels because, in a way, they are too big—or rather they cross-cut levels: they are transverse sections which do not include more than a small fraction of the phenomena at any given level..." (*op. cit.*, p. 19). It is more economical to identify levels in a neutral spatiotemporal fashion, however, and to go on to distinguish *domains* by coupling levels with perspectives (or sets of interests), as I have done in the text. That would enable us to avoid saying that levels are in some ways special cases of perspectives, in other ways larger than perspectives, and in yet other ways smaller.

Note also that Dudley Shapere has made some use of the notion of a scientific "domain" in a number of writings. But he does not incorporate interests in his domains; rather, he takes a domain to be a "body of information" to be investigated in a certain area or field, and contrasts it with the body of background information relevant to the domain—see *Reason and the Search for Knowledge* (Dordrecht: Reidel, 1984), Introduction.

The image of the scientific enterprise that emerges from these considerations is of a patchwork of overlapping theories or taxonomies, some of which inquire into different spatiotemporal levels, whereas others inquire into the same level relative to different interests. It is futile to look for a single system of categories that uniquely classifies a certain set of scientific phenomena. For any given spatiotemporal level, there does not seem to be a limit on the number of taxonomic systems that may categorize it, so uniqueness is not a promising construal of what makes a system of categories "natural." Moreover, systems of categories that categorize the same level relative to different interests, as well as those which categorize different levels, are not incommensurable rivals, for they can coexist comfortably in a single scientific account of the world. What is the upshot of this discussion for the notion of a natural kind? As I remarked in the opening paragraph, naturalness can be interpreted in various different ways.[21] But the particular notion of natural kind examined here—which pertains to a system of categories that cannot be crosscut by another system of natural categories—should have been rendered suspect by the above discussion. If equally legitimate categories can crosscut one another, then there is no hope for the dream of a single overarching taxonomic hierarchy for the whole of nature. Therefore, this notion of natural kind is not in tune with real ways of classifying things and ought to be rejected, though some other notion of natural kind may yet be articulated.

MUHAMMAD ALI KHALIDI

American University/Beirut

---

[21] Hacking has also pointed out on more than one occasion that there are many different notions of natural kinds—see, for example, "A Tradition of Natural Kinds," p. 122.

NOTES AND NEWS

The Editors regret the untimely death of Robert E. Butts, Professor of Philosophy at the University of Western Ontario, on March 26, 1997. He was 68 years old. Professor Butts earned his A.B. and A.M. from Syracuse University (1951 and 1952) and his doctorate from the University of Pennsylvania (1957). In 1963, he left Bucknell University, where he had taught since 1958, and came to Western where he served first as Head then as Chair of the Philosophy Department. In his capacity as Chair, Professor Butts oversaw the consolidation of the department into one of Canada's major centers for philosophy. From 1980-1990, he served as Editor of *Philosophy of Science* and was, since 1974, the general Editor of the *Western Ontario Series in the Philosophy of Science*. Among his many writings, he will be remembered best for his work on Immanuel Kant, *Kant's Double Government Methodology* (1984), and for his papers on William Whewell.

The XXth World Congress of Philosophy will be held in Boston on August 10-16, 1998. The following are among the confirmed invited speakers as of this writing: Henry Allison; William Alston; Karl-Otto Apel; Pierre Aubenque; Michael Ayers; George Bealer; Akeel Bilgrami; Paul Boghossian; Laurence BonJour; Michael Bratman; Dan Brock; Steven Cahn; Margaret Chatterjee; Debi Prasad Chattopadhyaya; Robert Coles; John Cottingham; Norman Daniels; Arthur C. Danto; Stephen Darwall; Donald Davidson; Martin Davies; Burton Dreben; Hubert Dreyfus; John Earman; Solomon Feferman; Richard Feldman; J. E. Fenstad; Arthur Fine; Kit Fine; John Martin Fischer; Robert Fogelin; Dan Garber; Bernard Gert; Gerd Gigenrenzer; Margaret Gilbert; Alvin Goldman; Adolf Grünbaum; Amy Guttman; Susan Haack; Bob Hale; Russell Hardin; Peter Hare; Gilbert Harman; Virginia Held; Risto Hilpinen; Ted Honderich; Christopher Hookway; Frances Kamm; Robert Kane; Jerrold Katz; Jaegwon Kim; Keith Lehrer; E. Jonathan Lowe; William Lycan; Alasdair MacIntyre; Gareth B. Matthews; Thomas McCarthy; Ruth Millikan; Jitendra N. Mohanty; Stephen Neale; Kai Kielsen; Christopher Norris; Martha Nussbaum; Charles Parsons; John Passmore; Günter Patzig; John Perry; John Pollock; Zenon Pylyshyn; W. V. Quine; Philip Quinn; Richard Rorty; Stanley Rosen; Jay Rosenberg; David Rosenthal; Sandra Rosenthal; William Rowe; David-Hillel Ruben; Michael Ruse; Esa Saarinen; Niels-Eric Sahlin; Mark Sainsbury; Israel Scheffler; Stephen Schiffer; John Searle; Sydney Shoemaker; Harvey Siegel; John Silber; Peter Simons; Walter Sinnott-Armstrong; Lawrence Sklar; Robert Sleigh; Michael Slote; Ninian Smart; Barry Smith; Peter Smith; Elliott Sober; Robert Sokolowski; Mark Steiner; Stephen Stich; Richard Swinburne; William Tait; Raimo Tuomela; James Van Cleve; Bastiaan van Fraassen; Peter Van Inwagen; G. H. Von Wright; Merrold Westphal; Howard Wettstein; Morton White; David Wiggins; Michael Williams; Timothy Williamson; Margaret Wilson; Kenneth Winkler; John Wippel; Paul Woodruff; Edith Wyschogrod; Linda Zagzebski. Further information and on-line registration are available at the web site http://www.bu.edu/WCP/ or by writing to: American Organizing Committee, Inc., 745 Commonwealth Avenue, Boston, MA 02215 (or to paideia@bu.edu).

THE JOURNAL OF PHILOSOPHY
VOLUME XCV, NO. 2, FEBRUARY 1998

COLLAPSE OF THE NEW WAVE*

There is a movement in the philosophy of mind recently dubbed "the new wave," one built upon "a novel general account of theory reduction in science."[1] Advocates hail Paul Churchland and Clifford Hooker,[2] claiming their model enjoys a number of theoretical advantages over its competitors. While discounting these advertised virtues, I argue that the new-wave model has but one genuinely novel element, and it should be rejected. Moreover, once its ontological consequences are duly noted, the model collapses into the classical theory developed within the logical empiricist tradition, so that it still falls prey to the standard antireductionist argument. The new wave is thus powerless to erode antireductionist sentiment and destined to recede into the distant past, leaving the conceptual landscape largely unchanged.

I. THE MODEL

New-wave reduction involves three theories: the basic reducing theory T_B, the original reduced theory T_R, and a corrected image T_R^*. With an eye to ideal cases, Paul Churchland describes it thus:

> [A] successful reduction ideally has the outcome that, under the term mapping effected by the correspondence rules, the central principles of

---

* My thanks to Charles Carr, Jerry Fodor, Thomas Grimes, Terence Horgan, Jaegwon Kim, John Post, and Kent Staley for their helpful comments on an earlier draft of this paper. I owe John Bickle a special debt of gratitude for many discussions and friendly disagreements about the new wave.

[1] John Bickle, "New Wave Psychophysical Reduction and the Methodological Caveats," *Philosophy and Phenomenological Research*, LVI (March 1996): 57-78, here p. 57.

[2] The model and its virtues are outlined in Paul Churchland, *Scientific Realism and the Plasticity of Mind* (New York: Cambridge, 1979); and "Reduction, Qualia, and the Direct Introspection of Brain States," this JOURNAL, LXXXII, 1 (January 1985): 8-28; and Hooker, "Towards a General Theory of Reduction, Part I: Historical and Scientific Setting; Part II: Identity in Reduction; Part III: Cross-Categorical Reduction," *Dialogue*, XX (1981): 38-59, 201-36, 496-529.

0022-362X/98/9502/53–72

> T_R (those of semantic and systematic importance) are mapped onto general sentences of T_B that are *theorems* of T_B. Call the set of such sentences $T_R{}^*$. This set is the image of T_R within T_B.[3]

Emphasizing a key aspect that will loom large in later discussion, he stipulates that the corrected $T_R{}^*$ must be part of the basic theory, being "general sentences of T_B." He also refers to a "mapping" with "correspondence rules" between this base-level $T_R{}^*$ and the original reduced T_R. Given the nature of $T_R{}^*$, these rules no longer have a familiar classical function. As Churchland goes on to say:

> First, on the account given above it is not the reduced theory, T_R, that is deduced from the principles of T_B, as some other accounts have it. What is deduced from T_B is rather the set $T_R{}^*$, an equipotent image of T_R within the idiom of T_B. Second, it is important to appreciate that cross-theoretic identity claims, even if they are justly made, are not part of the reduction proper, and they are not essential to the function it performs (*ibid.*, p. 83).

So reduction remains deduction. But unlike the classical account, crosstheoretic identity claims are "not part of the reduction proper," since it is not the original T_R but rather $T_R{}^*$ that is deduced from T_B, an image specified in the basic vocabulary. Hence, intertheoretic rules are inessential to both what is deduced and how. Instead, their primary function is to show how $T_R{}^*$ can capture the explanatory role of T_R. As Paul Churchland puts it:

> The correspondence rules play no part whatever in the *deduction*. They show up only later, and not necessarily as material-mode statements, but as mere ordered pairs: <*Ax*,*Jx*>, <*Bx*,*Kx*>, <*Cx*,*Lx*>, <*Dx*,*Mx*>. Their function is to indicate which term substitutions in the image $T_R{}^*$ will yield the principles of T_R. The older theory, accordingly, is never deduced; it is just the target of a relevantly adequate *mimicry*.[4]

Hooker expresses the emerging picture more formally in this way:

> Within T_B construct an analog, $T_R{}^*$, of T_R under certain conditions C_R such that T_B and C_R entails $T_R{}^*$ and argue that the analog relation, AR, between T_R and $T_R{}^*$ warrants claiming (some kind of) reduction relation R, between T_R and T_B. Thus $(T_B \quad C_R \rightarrow T_R{}^*) \quad (T_R{}^* \text{ AR } T_R)$ warrants $(T_B \text{ R } T_R)$.[5]

---

[3] *Scientific Realism and the Plasticity of Mind*, p. 81, with a slight change in the variables, henceforth. It should be noted that Churchland no longer speaks of theories in terms of public-language sentences, but prefers connectionist phase spaces. More on this in section IX.

[4] "Reduction, Qualia, and the Direct Introspection of Brain States," p. 10.

[5] "Towards a General Theory of Reduction, Part I," p. 49.

Of course, the analogue relation is not always strong, and the terms of T_R will not always map smoothly onto the terms of $T_R{}^*$. In Paul Churchland's words: "we must be prepared to count reducibility as a matter of degree. Like translation, which may be faithful or lame, reduction may be smooth, or bumpy, or anywhere in between."[6] So there is a continuum of strong to weak analogies between reduced and reducing theories, one that carries important metaphysical consequences. In the smooth case, though the *de*duction does not contain crosstheoretic identity claims, they are justified nevertheless by the success of the overall *re*duction. Churchland is explicit:

> On this view, full-fledged *identity* statements are licensed by the comparative *smoothness* of the relevant reduction (i.e., the limiting assumptions or boundary conditions on T_B are not wildly counterfactual, all or most of T_R's principles find close analogues in $T_R{}^*$, etc.). This smoothness permits the comfortable assimilation of the old ontology with the new and thus allows the old theory to retain all or most of its ontological integrity. *It is smooth intertheoretic reductions that motivate and sustain statements of cross-theoretic identity, not the other way around.*[7]

On the other hand, when the analogy is weak and the particular case is not smooth, the ontology of the original theory is replaced, not reduced. As Patricia Churchland[8] says:

> The evolving unifications seen in science therefore encompass not only smooth reductions with cross-theoretic identifications but also rather 'bumpy' reductions where cross-theoretic identifications are problematic and involve revision of the old theory's concepts, and outright elimination with no cross-theoretic identifications at all (*ibid.*, p. 284).

Or, again, mindful of the metaphysical consequences, John Bickle[9] describes it thus:

> [T]his alternative account of intertheoretic reduction recognizes a spectrum of possible outcomes, and insists that the appropriate ontological consequences depend upon where on this spectrum a given case falls. We can lay out this spectrum (informally) as follows: at the left-most endpoint lie the perfectly smooth reduction pairs, where $T_R{}^*$ is the ex-

---

[6] *Scientific Realism and the Plasticity of Mind*, p. 84.

[7] "Reduction, Qualia, and the Direct Introspection of Brain States," p. 11. He then defines "property" reduction: "A property *F*, postulated by the older theory or conceptual framework T_R, is reduced to a property *G* in some new theory T_B just in case (1) T_B reduces T_R; (2) *F* and *G* are correspondence-rule paired in the reduction; and (3) the reduction is sufficiently smooth to sustain the ontology of T_R, and thus to sustain the identity claim, '*F*-ness = *G*-ness'" (*ibid.*).

[8] *Neurophilosophy: Toward a Unified Science of the Mind-Brain* (Cambridge: MIT, 1986).

[9] "Revisionary Physicalism," *Biology and Philosophy*, VII (1992): 411-30.

> actly equipotent isomorphic image of T_R .... At the right-most endpoint lie the extremely bumpy or replacement cases. And separating these two extremes is a continuous spectrum of cases, approximating more or less closely one of the two endpoints (*ibid.*, p. 417).

So, putting together the various points, new-wave reduction involves four main features:

(i) New-wave construction: the basic reducing T_B, not the original reduced T_R, supplies the conceptual resources for constructing the corrected T_R*.
(ii) New-wave deduction: the corrected T_R*, not the original reduced T_R, is deduced from the basic reducing T_B.
(iii) New-wave relation: there is a required analogical relation, not bridge laws, between the reduced T_R and the corrected T_R*.
(iv) New-wave continuum: there is a continuum of strong to weak analogies between the reduced T_R and the corrected T_R*, with the strong relations justifying retention and the weak relations justifying replacement of the ontology of T_R.

Viewed as a process, the account can be summarized in two stages: first, the *intratheoretic deduction*, which focuses on conditions (i) and (ii); second, the *intertheoretic mapping*, which focuses on conditions (iii) and (iv). Once the second stage is complete, meaning once the analogue relation between T_R* and T_R has been established, one can make a claim about the reduction of the original T_R by virtue of the deduction of its analogue T_R*.

II. AFFINITIES WITH LOCAL AND APPROXIMATE REDUCTION

Certainly, the new-wave features have no echo in the classical model. There were bridge laws, not mere analogies, between the reduced and reducing theories; those bridge laws played an essential role, not a corrected image, in the derivation of the original reduced theory; and even if a corrected image were implied, it need not have been constructed from the basic reducing theory.

Yet the new-wave model is not without precedent. For example, condition (ii) on new-wave deduction is implied by previous accounts of local or domain-specific reduction developed by Jaegwon Kim and David Lewis,[10] according to which a corrected T_R* is deduced from the basic reducing T_B, where the former is a domain-specific version of a more general and otherwise reductively recalcitrant T_R. Even closer in spirit and ideology, both conditions (ii) and (iii) are explicit in Kenneth Schaffner's[11] earlier model of approximate reduction. Specifically, like condition (ii), Schaffner stipulates "that T_R^* be derivable from T_B," a theory which "corrects" the original T_R; and like condition (iii), Schaffner stipulates that the "relation between T_R and T_R^* should be one of strong analogy," the predictions of T_R^* being "close" to those of the original (*op. cit.*, p. 144). Also, the new-wave continuum of condition (iv) is easily generated from Schaffner's account by loosening his strong analogy into a spectrum of strong to weak relations of the same kind. Moreover, on this point there will be no difference at all if one reserves the term 'reduction' for cases at the retentive end of continuum where, by new-wave lights, the analogy is strong (involving "close analogues" and "equipotent images"). Consonant with this well-established usage, Hooker preserves the distinction between 'reduction' and 'replacement' by applying the former exclusively to "cases which fall sufficiently far towards the retention end of the retention/replacement spectrum."[12] So, too, does Patricia Churchland, who employs the same distinction from the opposite direction, saying that disanalogous cases which require massive correction are "better described as having been displaced outright" (*op. cit.*, p. 311).[13]

---

[10] Kim, "Phenomenal Properties, Psychophysical Laws, and the Identity Theory," *Monist*, LVI (1972): 177-92, esp. p. 190; and his fullest treatment in "The Myth of Nonreductive Physicalism" and "Multiple Realization and the Metaphysics of Reduction," collected in *Supervenience and Mind* (New York: Cambridge, 1993), pp. 265-84, 309-35. Lewis, "Review of *Art, Mind, and Religion*," this JOURNAL, LXVI (January 1969): 22-35, esp. pp. 24-25; and his "Mad Pain and Martian Pain," in Ned Block, ed., *Readings in Philosophy of Psychology*, Volume 1 (Cambridge: Harvard, 1980), pp. 216-22.

So, on the matter of conditions (ii) through (iv), and being mindful that the topic is reduction and not any wider scientific practice that includes theory replacement, the Schaffner and Paul Churchland-Hooker models are virtually indistinguishable. Condition (i) on new-wave construction is the genuinely novel element: the basic reducing T_B and not the original reduced T_R must supply the conceptual resources for the corrected image T_R^*. Only with this constraint can the older Schaffner conditions appear in a new light, turning what might have been a corrected heterogeneous deduction of T_R^* from T_B into a purely intratheoretic one, or, contrawise, turning what might have been an intratheoretic analogy between T_R^* and T_R into a purely heterogeneous one.[14]

---

[11] "Approaches to Reduction," *Philosophy of Science*, XXXIV (June 1967): 137-47. See also William Wimsatt, "Reductionism, Levels of Organization, and the Mind-Body Problem," in G. Globus, G. Maxwell, I. Savodnik, eds., *Consciousness and the Brain: A Scientific and Philosophical Inquiry* (New York: Plenum, 1976), pp. 205-67, esp. p. 217.

[12] "Towards a General Theory of Reduction, Part I," p. 45. The rationale, for Hooker, is that reduction must still achieve some measure of explanatory and ontological unification.

[13] See also Patricia Churchland's remarks about what the reducing theory must explain (p. 283), as well as the distinction between "bumpy reduction" and "outright displacement" (p. 284), all of which presuppose the well-established usage in question. Yet see Paul Churchland (*Scientific Realism and the Plasticity of Mind*, p. 84), and Bickle (pp. 64-65), who describe *non*borderline, *non*retentive cases as "bumpy reductions." By standard usage, however, calling outright replacement a "bumpy reduction" is slightly perverse—like calling one's divorced status a "bumpy marriage."

II. A MATTER OF DISPLACEMENT

Proponents of the new wave claim their model exhibits a number of theoretical virtues. The first concerns displacement of the original theory. Paul Churchland says of the two-stage process:

> The *intra*-theoretic deduction (of T_R^* within T_B), and the *inter*-theoretic mapping (of T_R into T_R^*), constitute a fell-swoop demonstration that the older theory can be displaced wholesale by the new, without significant explanatory or predictive loss.[15]

More precisely, the new and corrected T_R^* has the resources to *mimic* the explanatory role of the original T_R in an ideally smooth case, and to *better* the explanatory role of T_R in a rough and disanalogous case, either way demonstrating the old's replaceability by the new. Surely, the point is intriguing, since, contrary to appearances, if Churchland is correct then the new-wave model does not yield a continuum from retention to replacement after all. No, it is *theory replacement across the board*.[16]

Yet there are problems. First, even if the original "theory" can always be displaced, this is not so for the theory's ontology; for in a smooth and strongly analogous case, the ontology of T_R is *retained*, since claims about identity are sustained. So, on this side of the spectrum, there is replacement only in a weak and attenuated sense, a mere exchange of labels with referents and roles largely preserved.

---

[14] I say 'might have been', since Schaffner does not *require* that T_R^* be constructed out of the resources of a higher-level T_R. Instead, his model is defined to include "homogeneous" cases (p. 144).

[15] "Reduction, Qualia, and the Direct Introspection of Brain States," p. 11. In *Scientific Realism and the Plasticity of Mind*, similar remarks were followed by the disclaimer: "Displacement, of course, need not actually take place, however much considerations of unity and simplicity might demand it. Familiarity, entrenchment, convenience, and continuity may together counsel a less puritan course" (p. 82). So it is a form of displacement *in principle*. Even so, Churchland implies that it *can* take place, and certainly "familiarity, entrenchment, convenience, and continuity" do not accord any explanatory privilege to the original reduced theory. On the contrary, in ideally smooth cases the corrected theory "will perform all the same predictive and explanatory functions" of the original (p. 82), and in less than ideal cases it will be "superior" (p. 83).

[16] Curiously, Bickle ("New Wave Psychophysical Reduction," pp. 64-67) chastises Jerry Fodor for claiming that reduction will eliminate (that is, *displace*) the special science theory, failing to note that Paul Churchland made this an explicit platform of the new-wave model. Nor is Bickle's plea for the preservation of special sciences based upon any essential explanatory function that would be lost in the displacement. Instead, like Churchland's concession to the impure demands of "familiarity,

Second, even this weak sense of replacement appears suspect. Even if we take theories as sets of sentences, for example, T_R should be individuated in such a way as to contain all the sentences adequate to express the objects and properties in its domain. But this will include T_R^* in ideally smooth cases, since, *ex hypothesi*, T_R^* is an equipotent image that preserves the ontology of the original. Indeed, Hooker indicates that T_R^* *is* the original T_R in cases of "perfect retention," which, either way, makes replacement impossible.[17] (It follows, moreover, that there is no uniform theory of reduction, since, at this endpoint of retention, T_R^* ($= T_R$) is not specified within T_B—assuming T_R is not—contrary to what the new wave generally requires.)

Third, and ideal cases to one side, some[18] argue that replacement only makes sense in the intralevel case where theories compete for the same logical space and target the same explanandum phenomena. If so, a higher-level T_R could not be eliminated by a purely base-level T_R^*. The novel new-wave constraint would thus show itself not a help but a hindrance.

Fourth, it is no accident that Paul Churchland refers to a "wholesale" replacement of T_R, since the corrected T_R^* must be specified entirely within T_B, excluding any contribution from the original T_R.[19] But, on the contrary, theories are seldom if ever replaced whole and entire. To cite

---

entrenchment, convenience, and continuity" (see footnote 15), Bickle is content to say only that: "Reducing theories are typically more general in scope than reduced ones, and so are often much more unwieldy for use in actual scientific practice" (pp. 66-67). There remains, then, a substantive disagreement: the new wave believes that sciences like psychology *can* be displaced without loss of explanatory power, retained only for the sake of convenience, while antireductionists deny it.

[17] Specifically, he says: "where perfect retention fails $T_R^* \neq T_R$" ("Towards a General Theory of Reduction, Part II," p. 203). Note, too, though Hooker conjectures elsewhere that "the retention extreme of the retention/replacement continuum goes unoccupied" (Part I, p. 45), that is a contingent matter, depending upon how actual theories may fare. So the model itself does not guarantee displacement, neither in principle, nor at the extreme endpoint of retention where $T_R = T_R^*$.

[18] See Wimsatt, "Reductionism, Levels of Organization, and the Mind-Body Problem," pp. 215, 222-23; also Robert McCauley, "Intertheoretic Relations and the Future of Folk Psychology," *Philosophy of Science*, LIII (June 1986): 179-99; and his "Explanatory Pluralism and the Co-evolution of Theories in Science," in McCauley, ed., *The Churchlands and Their Critics* (Cambridge: Blackwell, 1996), pp. 17-47.

[19] That one must take the new-wave constraint in this exclusionary sense is clear; otherwise, the deduction of T_R^* from T_B would not be in*tra*theoretic in nature, that is, in derivations where elements of a higher-level T_R remain in T_R^*.

a well-known case, Copernican theory retained much of the Ptolemaic scheme, including all the heavenly bodies and the same celestial mechanics, only slightly fewer epicycles and eccentrics.[20] As the new wave insists for reduction, so, too, for replacement—it is a matter of degree.

Fifth, and finally, other models of reduction offer a similar benefit: for example, the views developed by Schaffner and Kim, since they also traffic in a corrected theory T_R* that can better the explanatory role of the original T_R, thus establishing T_R's displaceability.[21] In fact, they offer a more *convincing* displacement if that phenomenon most plausibly occurs when T_R and T_R* are intralevel competitors, or when the replacement is partial rather than wholesale; for unlike the new wave, these other views allow T_R* to be specified by, and thus retain, the conceptual resources of an original higher-level T_R.

IV. REDUCING THE FALSE BY THE TRUE

So consider a second virtue. Proponents of the new wave claim their model can account for the otherwise puzzling case where the original reduced theory is false. Compare the classical view. If T_R is deduced from T_B, and if T_R is false, then it seems to follow that T_B must be false. As Paul Churchland explains:

> Difficulties with this view begin to emerge with the observation that most reduced theories turn out to be, strictly speaking and in a variety of respects, *false*. (Real gases don't really obey $PV = \mu RT$, as in classical thermodynamics; the planets don't really move in ellipses, as in Keplerian astronomy; the acceleration of falling bodies isn't really uniform, as in Galilean dynamics, etc.) If reduction is *de*duction, modus tollens would thus require that the premises of the reducing theories (statistical thermodynamics in the first case, Newtonian dynamics in the second and third) be somehow false as well, contrary to their assumed truth.[22]

---

[20] This is Copernicus' considered view, found in the later books of the *De Revolutionibus*. See Thomas Kuhn, *The Copernican Revolution: Planetary Astronomy in the Development of Western Thought* (Cambridge: Harvard, 1957). Kent Staley has pointed out to me that the historical fact of partial replacement was emphasized long ago by William Whewell in his essay, "Of the Transformation of Hypotheses in the History of Science," *Transactions of the Cambridge Philosophical Society*, IX (1851): 139-47, reprinted in Robert E. Butts, ed., *William Whewell: Theory of Scientific Method* (Indianapolis: Hackett, 1989), pp. 251-62.

[21] Accordingly, I have argued that Kim's model is best viewed as a form of eliminativism via the original higher-level theory. See my "Species-Specific Properties and More Narrow Reductive Strategies," *Erkenntnis*, XXXVIII (May 1993): 303-21, esp. pp. 306-10.

[22] "Reduction, Qualia, and the Direct Introspection of Brain States," p. 9. Notice that Churchland comes dangerously close to a fallacy of equivocation: modus tollens needs "truth," Newtonian dynamics "assumed truth" (more precisely, "assumed truth *at the time proposed*," since it is now known to make false predictions when velocities near the speed of light, and so on).

On the new-wave model, however, one deduces a *corrected* version T_R* which can be sufficiently *unlike* the original and incorrect T_R so as to yield a difference in truth value. In Paul Churchland's words: "on the liberal account we are here embracing it is clear that a true theory may reduce a false one."[23]

But matters are far from clear. First, and again, there is nothing particularly "new wavish" about the present virtue, since it has nothing to do with the novel constraint on theory construction. Rather, it concerns the condition on *deduction*, that T_R* and not the incorrect T_R be derived from T_B. Accordingly, Schaffner's account yields the same result, as does Kim's and any other view that deduces a corrected, approximate, fragmented, or domain-specific counterpart whose truth value can differ from the original.

Even the classical model may yield the same result, since the original T_R is not deduced from T_B alone, but from T_B in conjunction with the pertinent connecting principles and boundary conditions. Any falsehood implied by the derivation could then be located in these other elements, not in the basic T_B (hypothetical connections and conditions still permit valid deduction, their usefulness being judged by the closeness of the adjoining domains, the amount of falsehood in the reduced theory, and the like).[24] Paul Churchland mentions this type of response, but adds:

> This defense will not deal with all cases of falsity, however, since in some cases the reduced theory is so radically false that some or all of its ontology must be rejected entirely, and the "correspondence rules" connecting that ontology to the newer ontology therefore display a problematic status.[25]

This is not convincing, however; for if the original theory is "radically false," then by the classical model and well-established usage, it is simply not a target for reduction. In the old parlance, theories of this kind are subject to "replacement," where this contrasts with reduction in its definitive role as the provider of ontological and explanatory unification. Put differently, since radically false theories do not even approximate the ontology of the supposedly true reducing theory, there will be no correspondence rules to affect the classical derivation and thus unify the two domains. Contra Paul Churchland's final remark, the rules do not then "display a problematic status." No, in radically false cases they simply do not exist! So the point remains: at least for moderately false cases, the classical derivation goes through, the false by the true, hand-in-hand with the necessary *façon de parler*.

---

[23] *Scientific Realism and the Plasticity of Mind*, p. 84. See also his "Reduction, Qualia, and the Direct Introspection of Brain States," pp. 9-10. Parenthetically, Fodor has pointed out to me in correspondence that whether the model allows T_R to be false while T_B is true depends entirely on the (unstated) constraints that govern the intertheoretic mapping. For example, if biconditional bridge laws are in the offing (as they will be at the retentive end of the continuum, see section VIII), then the same trouble affects the new-wave account.

[24] This only forces one to reconstrue the nature of the deduction as a kind of "transformation" versus a "sound argument" with true premises. See Thomas Nickles, "Two Concepts of Inter-theoretic Reduction," this JOURNAL, LXX, 7 (April 12, 1973): 181-201.

[25] "Reduction, Qualia, and the Direct Introspection of Brain States," pp. 9-10.

Finally, even granting that the classical theory should accord T_R and T_B the same truth value, this may not be an untoward result. Suppose a scientific theory is counted true only when all constituent propositions are true, false when at least one proposition is false (a trivial result on the assumption that a theory is defined as the logical conjunction of all constituent propositions). If so, then the reduced and reducing theories will have the same truth value: both will be false. That is, barring the philosopher's ideal science or God's omniscient perspective, every basic reducing theory is likely to contain at least one false proposition. Indeed, well-confirmed induction over past history will testify to the same, that basic reducing theories are false, strictly speaking, and including the very examples Paul Churchland cites, like the now outdated Newtonian dynamics reducing Keplerian and Galilean theories. Hence the desire to reduce the false by the true will go unfulfilled. Error is the way of the world.[illegible]

V. THE EXPLANATORY BURDEN OF CONNECTING PRINCIPLES

So consider a third supposed virtue. Bickle claims that, given the new-wave model, there is no need to explain the status of bridge laws between reduced and reducing theories. Referring to the deduction of T_R^* rather than the original T_R, Bickle says:

> This contrast has a key consequence: it eliminates what was a vexing problem for the orthodox empiricist account of reduction, namely that of specifying the logical and ontological status of the bridge laws, espe-

---

[illegible] Terry Horgan has suggested to me that, while my arguments are technically correct, the spirit of "reducing the false by the true" might be better served by saying that new-wave reduction can eliminate one *source* of error—falsity in the higher-level theory—even if there are other sources of error in the reducing theory. Yet Horgan notes that this advantage is not exclusive to new waveism, which is my central point. I also add deep reservations about the underlying epistemic principle; for the one domain is not inherently more error-ridden than the other. So, for example, it is not like the (already suspect) empiricist scruple whereby one chooses "fallible observation" over "fallible observation plus even more fallible commitment to unobservables." If there is no reason to favor low-level error, then impartiality would demand that one either eliminate all sources or retain both levels instrumentally without commitment to their truth.

> cially in reductions that imply or reflect significant ontological change (*op. cit.*, p. 58).

Bickle repeats the claim elsewhere, saying: "One advantage of the H-C [Hooker-Churchland] account is that it avoids having to specify the logical status of cross-theoretic identity statements, a problem that becomes especially pressing for the bridge-law account when a reduction entails significant correction to the reduced theory."[27]

Yet Bickle's remark that the problem is especially acute for cases that reflect "significant ontological change" echoes Paul Churchland's confusion about the status of correspondence rules when the original theory is radically false. On the classical view, significant ontological change calls for theory replacement, not reduction, meaning that there will be no system of bridge laws to affect the classical derivation, and hence no question about their logical or ontological status.[28] Moreover, when there is no significant change, then the status of the bridge laws is infamously clear—they are *identity* statements, the considered opinion came to be, an opinion originally offered to solve (among other things) the very problem at issue.[29] Surely, the new wave cannot complain, since their model is also committed to crosstheoretic identities as a consequence of relatively smooth cases at the retentive end of the reductive spectrum.

But the important point, for purposes of explaining bridge laws, is that the tables can be turned. A defender of the new wave inherits a weightier burden on grounds that she has *more* intertheoretic principles to explain, not just those implicated by ontologically retentive cases, but an intertheoretic mapping across the entire spectrum from smooth reduction to bumpy replacement. That is, the pertinent new-wave mapping also encompasses "correspondence rules" or "ordered pairs" in rough and disanalogous cases, their function, again, being "to indicate which term substitutions in the image T_R^* will yield [mimic or better?] the principles of T_R."[70] Nor is this intertheoretic mapping entirely unproblematic, seeing that, in nonretentive cases, T_B (and thus its subset T_R^*) are often *radically incommensurate* with respect to the original T_R. How, then, can radically incommensurate terms be mapped upon or substituted one for another? Is this an idealization, the classical *façon de parler*?[71] The alleged sins of the old are repeated tenfold by the new.

---

[27] "Mental Anomaly and the New Mind-Brain Reduction," *Philosophy of Science*, LIX (June 1992): 217-30, esp. p. 223. See also his "Multiple Realizability and Psychophysical Reduction," *Behavior and Philosophy*, XX (1992): 47-58, esp. p. 54.

[28] What Bickle could argue is that the status of bridge laws is problematic when there are "minor" ontological changes; for in that case reduction, not replacement, is still in the offing, yet the classical resources of strict identity via bridge laws might be incompatible with these differences. Yet, in my view, such arguments ultimately conflate differences in meaning with differences at the level of reference and ontology. In any case, the new wave is saddled with the same strict identities in ontologically retentive cases. See section VIII.

[29] See Lawrence Sklar, "Types of Inter-Theoretic Reduction," *British Journal for the Philosophy of Science*, XVIII (August 1967): 109-24; Robert Causey, *Unity of Science* (Dordrecht: Reidel, 1977), chapters 4 and 5; and a nice summary in Jaegwon Kim, *Philosophy of Mind* (Boulder: Westview, 1996), pp. 212-16.

VI. CO-EVOLUTION AND NEW-WAVE CONSTRUCTION

I have examined three central new-wave virtues and found them too common or too objectionable; another acclaimed virtue will be discussed shortly. But I now want to present some positive arguments against new-wave theory. The first draws upon William Wimsatt's observations concerning "co-evolutionary development" between reduced and reducing theories, a doctrine which members of the new wave wholeheartedly embrace. As Patricia Churchland describes it:

> [T]heories at distinct theoretical levels often co-evolve [citing Wimsatt], as each informs and corrects the other, and if a theory at one stage of its history cannot reduce a likely candidate at a higher level, it may grow and mature so that eventually it does succeed in the reductive goal. In the meantime the discoveries and problems of each theory may suggest modifications, developments, and experiments for the other, and thus the two evolve towards a reductive consummation (*op. cit.*, p. 264; also pp. 362-76).[72]

---

[70] "Reduction, Qualia, and the Direct Introspection of Brain States," p. 10. There is no question that Churchland means to include rough and disanalogous cases; for he cites the "radically false" cases that make any deduction of the original problematic (pp. 10-11), and says "what cases like these invite us to give up is the idea that what gets *de*duced in a *re*duction is the theory to be reduced" (p. 10). Hence he presents the intratheoretic deduction, and then the intertheoretic mapping with the pertinent remarks quoted in the text.

[71] Paul Churchland also misses the problem. He says "it is possible on this account for a theory to reduce even an *incommensurable* competitor," and discusses the reduction of classical mechanics (CM) by the special theory of relativity (STR) (*Scientific Realism and the Plasticity of Mind*, p. 85). The reduction is possible, he says, in spite of the difference in how mass is conceived, and so on, because "it is easy to deduce within STR...a highly convincing image of CM" (*ibid.*), which is to say, because of the first stage where one avoids deducing the incommensurable T_R. But this fails to address the incommensurable "mapping" or "substitution" in the second stage between that same allegedly incommensurable T_R and the corrected base-level counterpart T_R^*.

[72] The original source is Wimsatt, pp. 230-37. See also Hooker, "Towards a General Theory of Reduction, Part I," p. 48. Contra Patricia Churchland, however, there is no reason to assume the theories will always co-evolve toward "reductive consummation." They may drift further apart or settle into a stable equilibrium. See McCauley's "Explanatory Pluralism and the Co-evolution of Theories in Science."

As Hooker and Bickle also emphasize, this co-evolutionary feedback is not unidirectional, or exclusively from low to high levels.[33] There is a top-down flow of information and influence. For example, 'cistron' is a corrected image of the Mendelian gene (a term in T_R^*, and hence a term supposedly formulated within the idiom of T_B). Yet it was not created from molecular genetics (T_B) *ex nihilo*, but from the pressure of the original Mendelian theory (T_R) to find a structure with the function of a gene. So *co*-evolved terms within T_B or rather its subset T_R^* are by their very nature dually constrained by the rationales and conceptual resources grounded at both levels. In a word, they are theoretical hybrids, mirror images of the intertheoretic correspondence rules within classical reduction, differing from them only by the cover of a single term. The moral is that, because of the natural ebb and flow between levels of scientific language and scheme, the basic reducing theory becomes permeated with high-level concepts and concerns.

Now the problem is straightforward. The new-wave constraint on theory construction stipulates that the basic T_B and *not* the original T_R must supply the conceptual resources for constructing the corrected image T_R^*. Yet this seems flatly contradicted by the fact that, once co-evolution has run its natural course, T_R^* has become a mutual product of T_B and T_R. How, then, is T_R^* specified "within the idiom of T_B" in any meaningful sense that *excludes* T_R?

The answer is not readily apparent. But it is clear what form the answer must take, namely, the new-wave advocate must discount the historical contribution made by the upper-level theory. But such a move will not register any deep semantic fact about the terms in question if historical properties carry some weight, as they do (directly or indirectly) by considerations about the social surround, remote baptismal groundings, linguistic divisions of labor, externalism about conventions, and naturally selected biofunctions.[34]

---

[33] Hooker, "Towards a General Theory of Reduction, Part III," pp. 513-14; Bickle, "New Wave Psychophysical Reduction," p. 76.

[34] Ludwig Wittgenstein, *Philosophical Investigations*, G.E.M. Anscombe, trans. (New York: Macmillan, 1953); Saul Kripke, *Naming and Necessity* (Cambridge: Harvard, 1980); Hilary Putnam, "The Meaning of 'Meaning'," in *Mind, Language and Reality: Philosophical Papers*, Volume 2 (New York: Cambridge, 1975), pp. 215-71; Tyler Burge, "Individualism and the Mental," in P.A. French, T. Uehling, Jr., and H.K. Wettstein, eds. *Midwest Studies in Philosophy*, Volume 4 (Minneapolis: Minnesota UP, 1979), pp. 73-121; and Ruth Garrett Millikan, *Language, Thought, and Other Biological Categories* (Cambridge: MIT, 1984). In fact, aside from current conceptual role, there is probably just one leading semantic theory that ignores historical factors—a covariational theory that depends upon counterfactual relations and not actual causes. See Fodor, *The Elm and The Expert: Mentalese and Its Semantics* (Cambridge: MIT, 1994), pp. 115-19. But even this is no ally

Moreover, the new wave cannot appeal to conceptual or inferential role as a way of partitioning off the contribution of the original upper-level theory,[23] since the model guarantees current inferential links between T_R^* and T_R, those secured by the intertheoretic mapping in the second stage of the new-wave reductive process, both inferences that yield identity and substitutions that indicate mimicry. A special class of these inferential links will block the otherwise quite natural suggestion that T_R^* counts within the idiom of T_B, exclusively, on grounds that it can be *deduced* from T_B. Likewise, T_R^* can be deduced from T_R in cases of perfect retention, since they are for all intents and purposes the very same theory. Similarly, T_R^* can be deduced from T_R in less than perfect but still strongly analogous cases, only now in conjunction with more generous boundary conditions, hypothetical assumptions, and convenient fictions (remember how the classical theory can deduce the false by the true).

VII. HISTORY, METHODOLOGY, AND THE INITIAL COLLAPSE

Suppose there is a way to reconcile new-wave theory construction with co-evolutionary development, which is to say, some plausible account of term-to-theory individuation which grants T_B sole propriety over T_R^*. Even so, the distinction between constructing T_R^* out of T_B versus T_R can only appear superficial from the perspective of the actual historical development of the sciences, seeing that crosstheoretic evolution guarantees a healthy interplay between the two levels. More important, new-wave theory construction can only appear unduly restrictive from the vantage point of reductionist ideology and methodology, since it rules out strategies that may require aid from above, from the level of concepts in T_R.

Consider local reduction, whereby a corrected T_R^* can be generated out of an original and more general T_R either by "relativizing"

---

to the present partitioning strategy mentioned in the text, since, for all cases at the *retentive* end of the new-wave continuum, the properties counterfactually related to T_R and T_R^* will be the same, and thus the terms will have identical semantic content.

[23] See Gilbert Harman, "Meaning and Semantics," in Milton Munitz and Peter Unger, eds., *Semantics and Philosophy* (New York: University Press, 1974), pp. 1-16; and Hartry H. Field, "Logic, Meaning, and Conceptual Role," this JOURNAL, LXXIV, 7 (July 1977): 379-409. Note that Paul Churchland has been committed to something like this account when speaking with the vulgar, for example, in *Scientific Realism and the Plasticity of Mind*, pp. 52-80. For his later state-space semantics, see "Some Reductive Strategies for Cognitive Neurobiology," in *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science* (Cambridge: MIT, 1989), pp. 77-110. For detractors, see Fodor and Ernest Lepore, *Holism: A Shopper's Guide* (Cambridge: Blackwell, 1992), and their "Paul Churchland and State Space Semantics," in *The Churchlands and Their Critics*, pp. 145-62.

the original types to a restricted domain or by viewing those general types as "fragmenting" into narrower ones.[36] Either way, it is an operation upon *original* higher-level types. The new concepts thus owe their existence to the old, if not entirely, at least in conjunction with familiar methodological pressures from the basic reducing theory (human pain is a modification of pain, constructed directly out of general psychology, but with an eye to finding some stable nonvariable structure within neuroscience).

The upshot is this: on the worst-case scenario, new-wave construction is flatly contradicted by co-evolutionary facts; on the best-case scenario, it is historically shallow and methodologically restrictive. Sound reason, therefore, enjoins that it must be rejected. Once rejected, the new-wave picture takes on a familiar look; for the corrected T_R^* may now be specified within the idiom of T_R, making the deduction of T_R^* intertheoretic in nature, and bringing in train all the hybrid correspondence rules that a corrected classical derivation must provide. More pointedly, since the constraint on theory construction is the only genuinely novel element in the new-wave model, and the three remaining conditions are either directly implied or easily generated from Schaffner's account, then the entire project will collapse into its predecessor's. 1980s new-wave reduction and 1960s revolutionary approximation are one.

VIII. BRIDGE LAWS AND THE FINAL COLLAPSE

The most important of all virtues is surely the capacity of the model to resist antireductionist criticism. To that end, all defenders of the new wave unite in holding that their view enables them to counter arguments that originally led to the demise of the classical theory. Without this, all would be for naught. Qua reductionism, the entire project would fail.

So consider those arguments which reject psychological and function-to-structure bridge laws, chiefly Hilary Putnam and Jerry Fodor's multiple realizability argument, and also Donald Davidson's[37] normativity argument, since they brought about widespread antireductionist sentiment with regard to psychology and the functionally specified domains of biology.

---

[36] See Causey, *Unity of Science*, pp. 147-49; Kim, "The Myth of Nonreductive Physicalism," p. 273.

[37] Putnam, "The Nature of Mental States," in *Mind, Language and Reality*, pp. 429-40; Fodor, "Special Sciences" in *Representations* (Cambridge: MIT, 1981), pp. 127-45; and my summary in "Multiple Realization," *The Encyclopedia of Philosophy*, Supplemental Volume (New York: Macmillan, 1996), pp. 365-66. For the normativity argument, see Davidson's "Replies," in B. Vermazen and M. Hintikka, eds., *Essays on Davidson: Actions and Events* (New York: Oxford, 1985), esp. pp. 245, 249; and Kim, "Psychophysical Laws," in E. Lepore and B. McLaughlin, eds., *Actions and Events: Perspectives on the Philosophy of Donald Davidson* (Cambridge: Blackwell, 1985), pp. 369-86.

Now observe the following new-wave strategy. Responding to Davidsonian arguments, Bickle claims: "the impossibility of psychophysical laws is irrelevant to the new thesis of mind-brain reductionism and the novel account of intertheoretic reduction underwriting it."[38] Why? Bickle asks and then answers the question:

> Having now presented the account of intertheoretic reduction adopted by proponents of the new mind-brain reductionism, what can we conclude about the Davidsonian objection based upon the impossibility of psychophysical laws? It fails, and quite conclusively. For the lack of cross-theoretic laws is of no consequence to whether an H-C reduction is possible, *since an H-C reduction nowhere requires bridge laws* (*ibid.*, p. 224; italics added).[39]

Generalizing, then, whether it be Davidson's normativity considerations or Putnam and Fodor's more influential multiple realizability argument, all fail because they challenge reductionism on the point of bridge laws between the reduced and reducing theories. This is a palpable hit when directed against the classical account, but one that falls wide of the mark when directed against the new wave, for "H-C reduction nowhere requires bridge laws."

But this is mistaken; for new-wave reduction includes an intertheoretic mapping in its second stage, which, when carried out for strongly analogous cases, justifies crosstheoretic identities. As Paul Churchland says: "a successful reduction of the ideal sort described provides an excellent reason for asserting the relevant cross-theoretic identities, the best reason one can have."[40] Yet property identity guarantees nomic coextension. So bridge laws exist within the new-wave account, being directly implied by retentive cases.

Nor will it do, as a response, to insist on a distinction between the "reduction proper" versus its "consequences," confining bridge laws to the latter.[41] Consequences are consequences, and to deny them is like a smuggler caught in the act whose only defense is: "I meant there was no contraband *on my person*!" Declared or no, up front or trailing behind in tow, the goods are there; and so the problem is clear:

---

[38] Bickle, "Mental Anomaly and the New Mind-Brain Reduction," p. 218.

[39] See also his "New Wave Psychophysical Reduction and the Methodological Caveats," pp. 58-59. Accordingly, this is a basic theme in his *Psychoneural Reduction: The New Wave* (Cambridge: MIT, forthcoming).

[40] *Scientific Realism and the Plasticity of Mind*, p. 83; also "Reduction, Qualia, and the Direct Introspection of Brain States," p. 11; Hooker, "Towards a General Theory of Reduction, Part I," pp. 45-46, and Part II.

[41] Bickle has suggested something like this in correspondence, telling me: "I only acknowledge [new-wave conditions] (i) and (ii) as part of the intertheoretic reduction relation proper. (iii) and (iv) are attempts to relate intertheoretic results to ontological conclusions." Granted, Paul Churchland also speaks of a "reduction

(i) If a case falls at the retentive end of the new-wave continuum, then crosstheoretic property identities exist between reduced and reducing theories.
(ii) If crosstheoretic property identities exist between reduced and reducing theories, then biconditional bridge laws exist between reduced and reducing theories.
(iii) Therefore, if a case falls at the retentive end of the new-wave continuum, then biconditional bridge laws exist between reduced and reducing theories.

The conclusion is inescapable: the retentive end of the new-wave spectrum will collapse into the classical account of reduction, replete with its intertheoretic bridge laws, and subject to all the slings and arrows of antireductionist argument against them. If some of these arguments are sound, as many believe, then the new wave is forced into a familiar position;[42] that is, to return full circle:

(iv) It is *not* the case that biconditional bridge laws exist between intentional psychology or functionally construed theories in biology and more basic physical theories.
(v) Therefore, it is *not* the case that intentional psychology or functionally construed theories in biology will fall at the retentive end of the new-wave continuum.

There is no type reduction. Rather, a defender of the new wave must locate all such theories at the replacement end of the new-wave

---

proper." Still, he did include the intertheoretic mapping and its ontological conclusions in the reduction (*Scientific Realism and the Plasticity of Mind*, p. 81; "Reduction, Qualia, and the Direct Introspection of Brain States," p. 11; and so on). Indeed, he counts "reducibility as a matter of degree," which can only mean the intertheoretic relations of conditions (iii) and (iv), surely not the intratheoretic deduction of (i) and (ii). Regardless, if we have a reduction of the original T_R at all, it must be tied to the relation between T_R and its image T_R* which conditions (iii) and (iv) address; otherwise, the deduction of T_R* from T_B would be completely irrelevant to the status of T_R.

[42] For defense of the Putnam-Fodor genre of argument, see my "On Physical Multiple Realization," *Pacific Philosophical Quarterly*, LXX (September 1989): 212-24. See also Terence Horgan, "Nonreductive Materialism and the Explanatory Autonomy of Psychology," in S. Wagner and R. Warner, eds., *Naturalism: A Critical Appraisal* (Notre Dame: University Press, 1993), pp. 295-320; and "Multiple Reference, Multiple Realization, and the Reduction of Mind," in F. Siebert and B. Preyer, eds., *Reality and Humean Supervenience: Essays on the Philosophy of David Lewis* (forthcoming). For the biological case, see Harold Kincaid, "Molecular Biology and the Unity of Science," *Philosophy of Science*, LVII (December 1990): 575-93.

continuum.[43] Surprisingly, then, matters remain much as they were before the new wave arrived upon the shore. In the absence of some other account, one must either accept antireductionism or embrace type eliminativism.

IX. CONCLUDING POSTSCRIPT

The astute reader will have gathered that the new-wave model has collapsed on several fronts. By rejecting its constraint on theory construction, the model collapses into approximate reduction. By observing bridge laws implied by its intertheoretic mapping conjoined with their ontological consequences, the retentive end of the model collapses into classical reduction. Finally, by accepting standard arguments against those bridge laws, the retentive end of the model collapses, period.

Let me add, in conclusion, a few remarks about certain newer, new-wave proposals that might be deemed relevant. In particular, Paul Churchland[44] no longer speaks of theories in terms of sentences or sets thereof, but rather as vectors through connectionist phase space. Bickle[45] has recently moved to a nonsentential account of theories understood in terms of model-theoretic structures. Therefore, one might claim that these *non*sentential accounts will enable the new wave to avoid intertheoretic correspondence rules and thus burn the bridge with classical reduction.[46] But not so; traditional correspondence rules are not avoided, only relocated.

---

[43] Of course, this is not to deny other cases of smooth reduction where type identity is preserved—for example, physical optics, Kepler's laws of planetary motion, or a domain-restricted reduction thermodynamics. Also, my remarks concern the retention versus replacement of *types*. I have not addressed Hooker's function-to-structure *token* reductions ("Towards a General Theory of Reduction, Part III," pp. 504-05; also Bickle, "Multiple Realizability and Psychophysical Reduction," pp. 55-56; and his *Psychoneural Reduction*). I feel there is no pressing need to address this view here, since the historical debate over reduction has always concerned types and not tokens, the *anti*reductionists being the token physicalists.

[44] The deep reason is that Paul Churchland has moved from a deductive nomological account of explanation to a psychological theory of prototype understanding grounded in the neurocomputational framework. See his "On the Nature of Theories: A Neurocomputational Perspective," and "On the Nature of Explanation: A PDP Approach," in his *A Neurocomputational Perspective*, pp. 153-56, 157-230. Similarly, Hooker has moved to a "regulatory systems" view of science which complements the connectionist paradigm. See C.A. Hooker, H.B. Penfold, and R.J. Evans, "Control, Connectionism, and Cognition: Toward a New Regulatory Paradigm," *British Journal for the Philosophy of Science*, XLIII (1992): 517-36; and Hooker, *Reason, Regulation, and Realism: Toward a Regulatory Systems Theory of Reason and Evolutionary Epistemology* (Albany: SUNY, 1995).

[45] See his *Psychoneural Reduction*; but compare Paul Churchland's less than enthusiastic view of the model-theoretic approach in "On the Nature of Theories: A Neurocomputational Perspective," pp. 157-58.

[46] So Bickle tells me that "no bridge laws or correspondence rules exist anywhere in my [forthcoming *Psychoneural Reduction*, not the earlier] new wave account of in-

No one wishes to deny the existence of public-language sentences, or their use by scientists when announcing, demonstrating, and publishing their theories (whatever the ultimate nature of theories might be). Accordingly, William Bechtel[47] proposes what amounts to a two-factor theory, supplementing internal neurocomputation with publicly manifested deductions and diagrams and "various external representational systems" (*ibid.*, p. 126). Paul Churchland,[48] too, accommodates these plain facts of institutional science, though he now accords them a secondary role within a broader scheme of theoretical activity that encompasses animals, protohumans, children, and nonscientific lay persons.

In fact, the public domain of sententialist epistemology seems absolutely necessary for the preservation of the new-wave model. Specifically, and in light of Paul Churchland's proposal, the condition that T_R^* be deduced from T_B cannot be internalized within neurocomputational-level systems, since prototype activation in a connectionist net does not involve rule-based deduction (ironically, traditional syntactic-driven architectures fare better on this score).

Worse still, there seems to be no plausible neurocomputational corollary for the one novel new-wave constraint that T_R^* be constructed out of the idiom of T_B rather than the original T_R. Would it mean that, as a matter of brute fact, all human brains are so constituted that T_R^*'s activation vector is always a subsection of T_B's, and that T_R's activation vector never overlaps either on T_B or its subsection T_R^*? That should depend entirely upon how the theories become calibrated within each individual, which is to say, upon the vicissitudes of the several input histories cum variable connections and weights.

So, belabored but important, the public language of science remains. Yet, also, advocates of the new wave have not retreated from the position that their model delivers important ontological conse-

---

tertheoretic reduction. How could they? Bridge laws and correspondence rules are sentential items, and my account of theory structure and intertheoretic relations nowhere employs sentences. Theories are sets of models with a certain set-theoretic structure." Bickle is right. If narrowly focused on the models and their set-theoretic conditions, then no sentences can be found. But, as I argue here, taking a wider view of scientists and new wavers discoursing about the nonsentential items, coupled with new-wave ontological commitments for those items in retentive cases, reveals an old and familiar, that is, classical story.

[47] "What Should a Connectionist Philosophy of Science Look Like?" in *The Churchlands and Their Critics*, pp. 121-44.

[48] "Bechtel on the Proper Form of a Connectionist Philosophy of Science," in McCauley, ed., *The Churchlands and Their Critics*, pp. 265-70, esp. pp. 266-67.

quences, namely, property identities in strongly analogous cases. Hence the problem surfaces again: property identities conjoined with the public language expressions of the hypothesized nonsentential vector spaces/sets of models T_R and T_B (or its subset $T_R{}^*$) will yield nomic coextension for the predicates used in those public language expressions. The result is a set of intertheoretic bridge laws, and such is the classical model, in full sententialist dress. From this newest perspective, old bridging principles emerge at the metalevel, within public-language descriptions which all new wavers employ and to which all new wavers are unavoidably committed (whereof one must speak, one cannot remain silent).

To end with a fitting metaphor, the "collapse of the wave function" in quantum mechanics refers to the fact that a particular wave can be treated in a distinctly un-wave-like fashion, in terms of classical physics as a particle with specific location and velocity. Similarly, here the "collapse of the wave function" refers to the fact that a particular new wave can be treated in a distinctly un-wave-like fashion, in terms of classical reduction with bridge laws between reduced and reducing theories.

RONALD P. ENDICOTT

Arkansas State University

REVIEW ESSAYS

REWORKING SANDEL'S REPUBLICANISM*

Democracy's Discontent: America in Search of a Public Philosophy. MICHAEL SANDEL. Cambridge: Harvard University Press, 1996. 417 p. Cloth $24.95.

In *Democracy's Discontent*, Michael Sandel describes "America in search of a public philosophy," as the book's subtitle puts it. He documents the commitment to liberalism that he finds in the public life of the United States today, and he sets up a contrast between this philosophy and the older American commitment to republicanism.

The aim of the book is to rekindle enthusiasm for republicanism and to dampen the contemporary ardor for liberalism: for liberalism, that is, in the broad sense in which liberals may be conservative as well as progressive (5). The book attempts to achieve this end by taking us through developments in constitutional law and political economy where we can see liberal ideals taking over from republican ones and where we can gain a palpable sense of the loss involved in the retreat from republicanism.

I find Sandel's claims engaging—in particular, for the wealth of detail with which he documents them—and, on the whole, congenial. But in one crucial respect they leave me unhappy. They are worryingly indeterminate about the precise nature of America's lost republican ideals, about what those ideals would require of us as citizens, and about where they would lead governmental policy. Whenever the book focuses on such matters, it fades out before achieving a sharp level of resolution.

I raise this complaint in a spirit of construction, however, not despair. It turns out that, if Sandel will only go along with the characterization of republicanism which I have defended elsewhere[1]—and which, in broad outline, Quentin Skinner[2] pioneered—then he can escape such charges of indeterminacy and still uphold most of the themes that he parades. If he would allow us to rework his narrative a little, restructuring it around this account of republicanism, he would be able to tell a story that has a much greater claim on our attention—so, at any rate, I shall argue.

---

* My thanks, for comments, to Quentin Skinner and James Tully.

[1] *Republicanism: A Theory of Freedom and Government* (New York: Oxford, 1997).

[2] "The Paradoxes of Political Liberty," *The Tanner Lectures on Human Values*, Volume XV (Salt Lake City: Utah UP, 1985), pp. 227-50.

My argument is in three sections: (I) I summarize the salient points in Sandel's original narrative, identifying the indeterminacies that worry me; (II) I present the elements in my account of republicanism which are relevant to that narrative; and (III) I rework the narrative around those elements, showing how the worrying indeterminacies are thereby removed.

I. SANDEL'S ORIGINAL NARRATIVE

Neglecting his own order of exposition, I think we can identify five salient themes in Sandel's book. These are:

(1) American constitutional law has become rights based in practice and therefore neutral in aspiration: it aspires to impose no particular values.
(2) This neutralism shows up also elsewhere: for example, in a preference-based pattern of economic policy making;
(3) and in an interest-based way of conceptualizing and organizing politics.
(4) In making a concern for civic virtue irrelevant, such developments contrast with older, republican ways of thinking.
(5) We would do well to adopt such ways of thinking in preference to contemporary liberalism.

(1) The first theme, bearing on a rights-based interpretation of constitutional law, is the main topic of the first half of Sandel's book. He thinks of the rights-based approach as one which abstracts away from particular accounts of what is good and valuable and which tries to argue, independently of any particular conception of the good, that certain rights are sacrosanct. They have a claim on state and society which is prior to the claim of any alleged good.

Sandel argues that such a concern for rights was not particularly prominent in discussions around the time of the American War of Independence and the constitutional convention. "Only in the final week of the convention, as the proposed constitution was being prepared for submission to Congress, did George Mason of Virginia rise to say he 'wished the plan had been prefaced with a Bill of Rights'" (33). Of course, it is true that the anti-Federalists used the absence of a bill of rights as an argument against the Constitution. But this did not come of a commitment to a rights-based way of thinking, according to Sandel. "In opposing the Constitution, they sought to limit national power, and they found in the bill of rights the most popular, though not necessarily the most effective, way of doing so" (35). When James Madison proposed a bill of rights, then—the proposal, somewhat amended, was adopted in 1791—he did not do so out of a personal commitment to a rights-based approach. "This reversal reflected shrewd political strategy and astute political science" (37).

It was only in the aftermath of the Civil War, according to Sandel's story, that the rights-based approach began to come into its own. The Thirteenth, Fourteenth, and Fifteenth Amendments to the Constitution were designed to ensure the emancipation and equal status of slaves. But the second of those measures had a further, incidental effect. "The Fourteenth Amendment also imposed on the states certain restrictions that would transform the role of the Supreme Court in protecting individual rights. It established that no state may 'abridge the privileges or immunities of citizens of the United States' or 'deprive any person of life, liberty, or property, without due process of law' or deny to any person 'the equal protection of the laws'" (39).

Sandel goes on to argue that this amendment empowered the Supreme Court to protect individual rights from state infringement and that, in doing so, it led to the appearance, over a number of generations, of a distinctively rights-based approach. Thus, when the famous Lochner-versus-New York judgment (1905) struck down a law prohibiting the employment of bakery employees for more than sixty hours per week, the Supreme Court used the words: "The general right to make a contract in relation to his business is part of the liberty of the individual protected by the Fourteenth Amendment" (41).

The Lochner approach was given up by the Supreme Court in the period of the New Deal. But the way of thinking that it encapsulated became ever more entrenched, according to Sandel's account. Perhaps there was no right of the kind alleged in the Lochner judgment; perhaps the will of the democratic majority should be given sway in such cases. But there were other rights that should be protected, and protected even against democratic majorities. Those rights enjoyed precisely the status of trumps that the Lochner Court had given to the right to make employment contracts.

Which rights had that trumping status? Well, the reason that the right to make employment contracts of any kind was not a protected right, according to critics of the Lochner Court like Oliver Wendell Holmes, was that its claim to be a right was predicated on the truth of a particular economic philosophy (43, 46). So the rights that should be protected naturally came to be depicted as rights of such a fundamental kind that their acceptance did not presuppose any particular conception of the good life or the good society. This approach was established after World War II, when the Supreme Court came to represent the constitution as a framework of rights that is neutral among different ends.

Sandel goes on to illustrate the neutralism—the no-value neutralism, as he views it—that he finds in judgments of the Supreme Court over the past few decades. He sees it in the tendency to think of religious liberty, for example, not as freedom of conscience in an area of agreed and special importance for human beings, but as the freedom to choose on matters of personal preference: on matters about the importance of which the Court has no opinion (70). He sees it in development of the doctrine that the content of a given sort of speech is irrelevant to the question of whether the Court ought to protect the freedom of that kind of speech: there is no distinction that the Court is entitled to make between low-value and high-value expression (73, 90). He sees it in the shift from a civic to a voluntarist way of defending privacy. Under the civic defense, privacy should be protected insofar as it fosters social institutions—say, the institution of marriage—that are of agreed importance; under the voluntarist defense, privacy should be protected insofar as it enables individuals to exercise their autonomy in a way that, however unimportant or even obnoxious it may seem to most of us, is consistent with a similar exercise of autonomy on the part of others (94-100).

(2) I have given a lot of attention to the first theme in Sandel's book—that American constitutional law has become rights based and neutralist—mainly because he himself devotes all of the first part of his book to it. His second theme, asserting the presence of a similar no-value neutralism in American economic thinking, emerges in the second part of the book. But it emerges only at the end of a detailed historical discussion of how thinking about political economy developed in the United States over the past two hundred years. When it appears, it does not receive the same sort of documentation he gives to the first.

Sandel's historical discussion of the development of thinking about political economy is the most riveting part of the book. He documents a continuing concern, from the period of the Constitution down to the New Deal, with the effect of economic arrangements on the citizenry. The fear throughout this period is that such arrangements can subjugate workers and render them incapable of achieving the independence and civic virtue that republicanism has always looked for in citizens.

Should there be widespread manufacturing industry in America—this, as distinct from an economy of independent farmers—given the subject status of industrial workers? No, thought Jefferson in the early days of the republic: "Dependence begets subservience and venality, suffocates the germ of virtue, and prepares fit tools for the designs of ambition" (144). But did not the Lowell industrial model of the Jacksonian era offer some hope? Not according to protestors in 1834 who objected to "that haughty, overbearing disposition, that purse-proud insolence" that they found in their employers (153). The position of industrial workers, as it was seen by artisan republicans, was often no better than that of southern slaves: they were "wage slaves" rather than "chattel slaves" but the dependence they suffered was nonetheless real for that (172-74).

Still, was there not a basis for consolation in the fact that industrial workers, as free laborers, could aspire to owning and working land of their own in an expanding America? There was, according to the proponents of the free-labor movement in the 1850s and 1860s. But, alas, the free soil rapidly ran out, and, by 1870, two thirds of productively engaged Americans were wage earners (183-84). The Knights of Labor articulated a widely felt discontent about the standing of wage earners but the failure of their own attempts to establish cooperatives meant that the discontent was unrelieved. As they saw it, workers could never become self-supporting and could not hope to enjoy the proper status of a citizen: "no man whose bread and that of his children are dependent on the will of any other man, or who has no interest in his work except to please an employer, fulfills these conditions; a farmer of his own land does fulfill them" (187).

The upshot, as Sandel tells his story, is that with no hope of a republican reconciliation with the nature and conditions of industrial employment, people generally moved over to a different conception of economic arrangements. They ceased to worry in the name of republican values about the effects of such arrangements. They consoled themselves with the thought—ultimately, the value-neutral thought—that, if workers voluntarily enter into employment relations then, however demeaning those relations may be, they presumably answer to the preferences of the workers. This neutralism in economic thinking reached its apogee in the rise of consumerism. The best economic arrangements are not those which satisfy some pregiven conception of the good, such as that which republicanism offers; according to the consumerist credo, they are the arrangements that best answer to the preferences that people reveal in their consumption choices, including their choice of job. "Instead of asking how to elevate or improve or restrain people's preferences, it asks how best—most fully, or fairly, or efficiently—to satisfy them" (225). The lesson is that "government should not form or revise, or for that matter even judge, the interests and ends its citizens espouse; rather, it should enable them to pursue these interests and ends, whatever they may be, consistent with a similar liberty for others" (267).

(3) So much for the no-value neutralism that Sandel finds in contemporary American thinking about constitutional law and the political economy. His third theme is very closely related, as I see it; indeed, it is so closely related that it barely receives separate mention. Not only is America neutralist in its attitude to constitutional law and the political economy, according to this third theme, it is also neutralist in its attitude to politics. There is no longer an assumption, as there used to be in republican days (5, 166), that politics is about the deliberative identification and cooperative pursuit of certain common goods: the plural reference to "common goods" is appropriate, because people may not agree on any unitary account of the common good (320). Rather, politics is conceptualized as a process of competitive bargaining among different interest groups, which involves no attempt to evaluate the propriety of any of the interests represented.

This theme is very prominent in the work of another self-described republican, Cass R. Sunstein,[3] and it is remarkable that Sandel makes no reference to this fact. But it is clear that he shares Sunstein's sense that interest-group politics represents a betrayal of American republican ideals, one that parallels the betrayal that he documents in rights-based constitutional law and in a preference-based political economy. Thus, he rightly castigates "modern-day interest-group pluralists" for invoking the name of Madison. "For Madison, the reason for admitting interests into the system was not to govern by them but to disempower them, to play them to a draw, so that disinterested statesmen might govern unhindered by them" (131).

(4) We come finally to the theme in regard to which I find fault with Sandel's book and believe that his narrative needs reworking. He counterpoints the neutralist tradition described in his first three themes with the republican approach that he takes—surely rightly—to have marked earlier stages, in particular the founding stage, of American public life. The main point of contrast is predictable. The public philosophy of today—"the assumptions about citizenship and freedom that inform our public life" (4)—eschews all values in favor of a no-value neutralism. The earlier public philosophy looked

---

[3] *After the Rights Revolution: Reconceiving the Regulatory State* (Cambridge: Harvard, 1990); *The Partial Constitution* (Cambridge: Harvard, 1993); *Democracy and the Problem of Free Speech* (New York: Free Press, 1993); "The Enduring Legacy of Republicanism," in S.E. Elkin and K.E. Soltan, eds., *A New Constitutionalism: Designing Political Institutions for a Good Society* (Chicago: University Press, 1993).

unashamedly for a certain sort of virtuous citizenry and insisted that social and political institutions should be designed so as to form people in that virtuous mould: it ascribed a committed, formative role to those institutions, not the role of a neutral referee (6).

At this point, we naturally wait on more: we want to be told exactly why the republican philosophy required a virtuous citizenry; exactly what sort of virtue it recommended; and exactly how the institutions of government were supposed to facilitate that civic virtue and advance republican ideals. We want to be told more, in other words, on three crucial issues: first, the nature of the republican ideals that mandate civic virtue; second, the nature of the virtue that they mandate; and third, the nature of the institutions and programs whereby that virtue is to be encouraged and those ideals promoted. Sandel does tell us something more about these things, though only in a series of short comments that are distributed throughout the book. But what he tells us falls well short of a satisfying account.

What, according to Sandel, are the central republican ideals? The ultimate ideal, we are told, is republican freedom or liberty as distinct from liberal freedom or liberty (4). So what is this republican freedom? In most of his utterances, Sandel suggests that he follows the "strong version of the republican ideal" (26), according to which being free essentially—he would say, "intrinsically"—involves "sharing in self-government" (9). "Republican political theory teaches that to be free is to share in governing a political community that controls its own fate" (274). He distinguishes this strong version from "more modest versions of the republican ideal," according to which the relation between self-government and freedom is instrumental rather than definitional. He appears to be well disposed to such versions of the ideal insofar as they agree with the strong version that, even if instrumental, the relationship between self-government and liberty is nonetheless "internal," but he does not explain what an internal connection involves (26). I shall return to this theme later, but for the moment I simply note that his own way of speaking clearly suggests that he thinks of liberty, in the "strong" republican manner, as being definitionally tied up with participation in government.

To move to the second issue, what sort of virtue does liberty in this sense mandate? What does the republican "conception of the good society" as a self-governing republic require of its members (25)? The answer given, at its most explicit, is this: "republican theory does not take people's existing preferences, whatever they may be, and try to satisfy them. It seeks instead to cultivate in citizens the qualities of character necessary to the common good of self-government. Insofar as certain dispositions, attachments, and commitments are essential to the realization of self-government, republican politics regards moral character as a public, not a private, concern" (25). What qualities of character are to be cultivated? Those which are necessary for people to be able to deliberate fruitfully about the common good and to help shape the society's destiny. We are told that they include "a knowledge of public affairs and also a sense of belonging, a concern for the whole, a moral bond with the community" (5).

So much for the nature of the republican ideal, according to Sandel, and the nature of the civic virtue it requires. What, finally, of the institutions and programs whereby such virtue is to be facilitated and the republican ideal promoted? The institutions are not to be exclusive of the nonmainstream, the nonmale, and the nonpropertied, in the manner of premodern republics (318-19); and they are not to be invasive and homogenizing, for Sandel's republican hero is Alexis de Tocqueville, not Jean-Jacques Rousseau. "Unlike Rousseau's unitary vision, the republican politics Tocqueville describes is more clamorous than consensual. It does not despise differentiation. Instead of collapsing the space between persons, it fills this space with public institutions that gather people together in various capacities, that both separate and relate them. These institutions include the townships, schools, religions, and virtue-sustaining occupations that form the 'character of mind' and 'habits of the heart' a democratic republic requires" (320-21).

These remarks on the sorts of republican institutions he seeks do not yet tell us in which direction Sandel thinks his ideal state will go in important matters of public policy. But he offers a number of comments, often in passing, that indicate the drift of his expectations. Thus, he is in favor of greater equality: "The republican tradition teaches that severe inequality undermines freedom by corrupting the character of both rich and poor and destroying the commonality necessary to self-government" (330). While he has sympathy for those communal conservatives who see a greater role for morals in public life (309), he clearly thinks that bringing issues of morality into politics will not necessarily turn back the clock on a range of social issues. He suggests that what is wrong with the liberal tolerance of homosexuality, for example, is not that it runs counter to the moralistic views of many and is indeed a form of tolerance, but that it does not directly challenge those moralistic views, as presumably a republican approach would do: "unless those views can be plausibly addressed, even a court ruling in their favor is unlikely to win for homosexuals more than a thin and fragile toleration" (107). More generally, he argues that going republican does not mean giving up on individual rights; it means providing them with a new foundation: "Rights can be defended on a number of grounds, including the grounds that respecting certain rights is a way of cultivating civic virtue or of encouraging among citizens certain worthy practices or beliefs or qualities of character" (290).

I am unhappy, as I said, with Sandel's comments on these three issues: the nature of the ideal that republicans espouse, the nature of the civic virtue it requires, and the nature of the public institutions and programs that it would support. Apart from leaving the connection between self-government and liberty underspecified—talk of an internal connection is just not good enough—he fails to explain how self-government is supposed to be implemented, especially in a society as complex and large as the United States today; and he fails to say anything about how self-government is going to be made proof against what Jefferson himself described as the tyranny of the majority. When he goes on to say that republican liberty and self-government require civic virtue in the citizenry, he remains studiously uncommitted on what the content of that virtue is supposed to be; we are referred, vaguely, to the qualities essential for participants in self-government. This worrying indeterminacy about republican freedom and republican virtue carries over into a similar, breezy vagueness about the institutions and programs that such freedom and virtue are going to support or, when examples are given, about why they are going to support them. How can Sandel be sure, for example, that the virtuous citizens of his self-governing republic will really be tolerant of homosexuality? We are never told.

(5) The fifth and final theme that I mentioned in describing Sandel's book is that we should give up the liberalism that constitutes America's current public philosophy and endorse instead the republicanism that emerges in his narrative. I shall not say much about this claim except to note that it is advanced, not by showing the merits of republicanism, but rather by displaying the weakness of the liberal alternative. In his opening chapter, Sandel considers three defenses of contemporary liberalism: a utilitarian defense; a Kantian defense; and a defense on minimal, pragmatic grounds. He argues against utilitarianism that it does not respect differences among persons; against Kantianism, that it does not enable us to vindicate particularistic, communal attachments; and against minimalism, that pragmatic resolutions are often just not available: moral and religious convictions cannot always be put aside, as in the abor-

tion debate, and putting them aside in other cases may erode "the moral and civic resources necessary to self-government" (23).

You cannot beat something with nothing and what worries me about this way of arguing for republicanism is that it is essentially negative. Perhaps that is as it has to be, given the vagueness that I have alleged in Sandel's characterization of republicanism. But it would have been much more satisfying had we been given a less elusive account of the approach and a more positive argument in its favor.

II. ELEMENTS FOR A REWORKING OF THE NARRATIVE

I am going to present three elements in my own account of republicanism and then show, in the section following, how Sandel's story can be reworked around those elements and how the problems raised for the story can be thereby removed or at least diminished. The first of my three elements bears on the definition of republican liberty, the second on the relationship between liberty in that sense and the institutions—including institutions of self-government—that support it, and the third on the relationship between those institutions and civic norms and virtues.

(1) The definition of republican liberty. When Sandel goes along with the view of republican liberty which holds that "to be free to share in governing a political community that controls its own fate" (202, 274), then he follows the neo-Athenian reading of republicanism popularized in recent times by Hannah Arendt; indeed, he refers to Aristotle and Arendt as the sources of the view (355). That reading associates republicanism with what Benjamin Constant,[4] early in the last century, called the liberty of the ancients and it lets liberalism claim the liberty of the moderns: this is liberty in the sense of being let alone, liberty as noninterference. Where liberalism would allow each person full control over her own domain of action, within limits dictated by the need for social coordination, republicanism on this account would allow each person partial control over the domain of everyone's action: the sort of control allegedly exercised by the citizens of classical Athens. Where liberal freedom consists in having nearly global control over a small domain, republican freedom would consist in having a small degree of control over a global domain: that of what everybody does.

Sandel allows that for Niccolò Machiavelli, as Skinner[5] has argued, most people pursue freedom in a sense that does not give them a share in government: "A small part of them wishes to be free in order to rule; but all the others, who are countless, wish freedom in order to live in security."[6] He happily concedes that Madison was primarily concerned to guard against "elective despotism" (36)—hardly the concern of someone who thought of freedom as sharing in self-government—and that he and the other framers of the Constitution pursued a "revision of classical republican assumptions" (130), as Sandel thinks of those assumptions. But when he suggests that the doctrines of Machiavelli and Madison are just variations on his own neo-Athenian themes, then I think he is quite wrong. What we find in Machiavelli and Madison—and what we will find in the long tradition of republicanism in the period in between[7]—is a distinct neo-Roman republicanism, Ciceronian rather than Aristotelian in inspiration. This is the account of republicanism, in particular of republican freedom, which I wish to introduce as the first element in a reworking of Sandel's story.

---

[4] *Constant: Political Writings*, B. Fontana, ed. (New York: Cambridge, 1988).

[5] "Machiavelli on the Maintenance of Liberty," *Politics*, XVIII (1983): 3-15.

The republican tradition from Machiavelli to Madison saw the Roman constitution through the eyes of Polybius and saw the practice of Roman politics through the eyes of Livy, Plutarch, Sallust, Tacitus, and, of course, Cicero himself. These republicans pictured Athens, rightly or wrongly, in the critical terms suggested by Polybius: as a ship without a captain, buffeted by the storms of popular opinion. They traced the Athenian problem to excessive reliance on pure democracy and saw the Roman republic, by contrast, as a constitution in which government was built on a democratic foundation but was better devised to guard against problems of faction and demagoguery and tyranny. The principal devices they celebrated in Rome were the dispersion of democratic power across different assemblies, adherence to a more or less strict rule of law, election to public office, limitation on the tenure of public office, rotation of offices among the citizenry, and the like. These devices provided the checks and balances whereby a republic might hope to constitute a government that was at once popular and stable.[8]

The key idea in this neo-Roman tradition is republican freedom, for it is the freedom ideal that provides the argument, ultimately, for recourse to such institutions. Republican freedom, however, is distinct both from liberty as noninterference and, Constant notwithstanding, from liberty as democratic participation.[9] It is a negative form of liberty, like the liberal ideal. But it requires something other than the absence of interference. It requires the absence of dependency upon the will of another and the absence of vulnerability to interference at the will of that other: it requires the absence of mastery or domination by any other. The antonym of freedom for the republican conception is not restraint as such but rather slavery and, more generally, any position of subjection. A person is free, and a person acts freely, just to the extent that she is not exposed, in the way a slave is exposed, to the arbitrary interference of another: to the sort of interference that only has to track the *arbitrium*—the will or judgment—of the interfering power.

---

[6] Machiavelli, *The Chief Works and Others*, Allan Gilbert, trans. (Durham, NC: Duke, 1965), p. 204.

[7] See M.N.S. Sellers, *American Republicanism: Roman Ideology in the United States Constitution* (New York: University Press, 1994); Skinner, *Liberty before Liberalism* (New York: Cambridge, 1997); and my *Republicanism*.

[8] See Sellers on the themes of this paragraph.

Freedom as nondomination differs in two striking ways from freedom as noninterference. The conceptions will agree that someone who is subject both to domination and to interference is unfree and that someone who is subject to neither domination nor interference is free. Where they differ is in the judgments made about those who are dominated but not actually interfered with—say, the subjects of a kindly master—and about those who are not dominated but are interfered with—say, the citizens who are bound by a regime of law that, as they themselves see it, tracks their common interests and ideas and is the contrary of arbitrary interference.

The sort of republicanism associated with the English commonwealthman tradition, ultimately the sort of republicanism shared by the American founders, was forthright on both of these themes. The commonwealthmen insisted, first of all, that being subject to another in the manner of a slave made a person unfree, even if the other was not disposed to interfere. Algernon Sydney[10] could write in the late 1600s, for example, that "he is a slave who serves the best and gentlest man in the world, as well as he who serves the worst" (*ibid.*, p. 441). In the century following, Richard Price[11] could add: "Individuals in private life, while held under the power of masters, cannot be denominated free, however equitably and kindly they may be treated" (*ibid.*, pp. 77-78).

The commonwealthmen also insisted, to move to the second point of contrast with liberalism, that insofar as a rule of law was genuinely nonarbitrary—insofar as it tracked the interests and ideas of citizens—to that extent it did not compromise their liberty. Not only did it protect people against a compromise of their liberty by others, it did not itself impose the sort of compromise that would have to be justified by such protection. John Locke,[11] the hero of eighteenth-century American republicans, is a good example of this attitude. He argues for a "freedom from Absolute, Arbitrary Power" as the essential thing, and the thing that marks the contrast with slavery. That leads him to represent law as in no way a restriction on freedom: "that ill deserves the Name of Confinement which serves to hedge us in only from Bogs and Precipices...the end of Law is not to abolish or restrain, but to preserve and enlarge Freedom" (*ibid.*, pp. 325, 348). Price again echoes this theme when he writes: "Just government...does not infringe liberty, but establishes it. It does not take away the rights of mankind but protect[s] and confirm[s] them" (*op. cit.*, p. 81).

---

[9] See my *Republicanism* and Skinner's *Liberty before Liberalism.*

[10] *Discourses concerning Government*, T.G. West, ed. (Indianapolis: Liberty Classics, 1990).

[11] *Political Writings*, D.O. Thomas, ed. (New York: Cambridge, 1991).

(2) Liberty and republican institutions. So much for the first of the three elements I want to introduce. The second bears on the relationship between liberty in the sense of nondomination and those public institutions through which such liberty is secured for the citizens of an ideal republic. The institutions that were taken to be required for the promotion of republican liberty are various, though they are almost all of a Roman provenance. They include institutions like a rule of law that governs everyone alike—even a monarch, if there is a monarchy; a dispersion of the powers given to those who make and run the law among many hands; a restriction on the tenure with which public office is held; a dispensation that allows for public discussion of important issues, for public deliberation over how to resolve them, and for a degree of public participation in their resolution; an arrangement for the appointment of certain officers by a system of voting or lottery or a mixture of both; and so on. The second element I want to introduce bears on how such institutions, assuming that there is a uniquely appropriate set for any social situation, are going to relate to the freedom of the citizens whom they are meant to support. Since the institutions were usually taken to include some measures of participatory government, this element bears in part on how participatory institutions are going to relate to the freedom of participating citizens.

Sandel is right that in the republican, even in the neo-Roman republican, way of thinking there is an intimate connection envisaged between the institutions that support freedom—between the constitution of the free society—and the freedom that they support. It is this connection, for example, that allows Charles de Secondat Montesquieu[13]—in many ways, a covert republican—to speak now of freedom in the constitution, now of freedom in the citizen. But the connection is not the definitional connection that relates freedom to participatory institutions when it is said, as Sandel says, that to be free is to share in the self-government of a community. It is also not the causal connection that relates freedom as noninterference to certain protective institutions: the connection whereby the presence of those institutions comes to inhibit potential offenders and thereby to have the causal effect of enhancing the freedom as noninterference of others. The connection between freedom as nondomination and relevant institutions is constitutive, as I have put it, and not definitional or causal (*op. cit.*, pp. 106-09).

---

[11] *Two Treatises of Government*, Peter Laslett, ed. (New York: Mentor, 1965).

Consider the connection between the antibodies that make someone immune to a certain disease and that immunity itself. The connection is not definitional, since immunity to the disease can be defined and understood without reference to the antibodies. But neither is the connection causal, for it is not as if the immunity is something distinct from the antibodies; being immune, for the person in question, comes to nothing more and nothing less than having those antibodies. The antibodies do not cause the immunity, as we say, they constitute it.

The relationship between people's freedom as nondomination and the institutions that would support that freedom in the ideal republic—assuming that an ideal republic is attainable—is going to be constitutive in the same sense. If freedom is nondomination, if freedom consists in the status of being more or less immune to arbitrary interference by others, then the institutions that ensure that immunity will relate to it neither in a definitional, nor in a causal way. Not in a definitional way, because the understanding and definition of nondomination will not require a reference to the precise institutions that work in the circumstances on hand; and not in a causal way, because the relationship between the institutions in question and the immunity to arbitrary interference that they confer will be exactly like the relationship between the antibodies in our earlier example and the immunity that they give against a certain disease. If freedom is noninterference, then freedom is causally brought about by suitably protective and empowering institutions: it is a result of the inhibiting effect of the protection and empowerment, now on this potential offender, now on that. If freedom is nondomination, then freedom just is the protected and empowered status—the invulnerability or nondomination—enjoyed in the presence of the institutions, and it exists prior to any potential offenders actually coming to be inhibited.

---

[13] *The Spirit of the Laws*, A.M. Cohler, B.C. Miller, and H.S. Stone, eds. and trans. (New York: Cambridge, 1989), esp. p. 187.

(3) Liberty and republican virtue. We have seen that republican freedom just is nondomination and that, if freedom in that sense is enjoyed by citizens, then it is constituted by the institutions that support it, not merely causally related to those institutions. The third element that I want to add is that, this being the case, we can also see a close connection between republican freedom and certain forms of civic virtue.

The connection comes of the fact, recognized throughout the republican tradition, that while the effective rule of law may confer freedom in the sense of nondomination, the laws themselves will only be effective in a society where they are supported by actively sustained norms. Thus, Machiavelli says: "just as good morals, if they are to be maintained, have need of the laws, so the laws, if they are to be observed, have need of good morals" (*op. cit.*, p. 241).[14] One argument for why the laws have need of good morals or customs—*buoni costumi*—comes in two stages. The first is that the laws will be effective in protecting and empowering each against others only insofar as the laws command universal allegiance, in particular allegiance among those in power. The second is that the laws will enjoy this command only insofar as people are normatively committed to the laws, not just committed to them for fear of the penalty attending noncompliance or underperformance.[15] The background thought here, supported by contemporary criminology, is that fear of penalty is not a very effective motive, on its own, in binding people to the law.

A second argument why laws have need of morals—have need of civic virtue or civility—is that, if the laws are effectively to track people's changing ideas and interests, and if they are to be guaranteed to remain nondominating, then there must be enough civic virtue in the society to ensure that those ideas and interests will be reliably represented in the forums of law making and the regime of law made to answer to them. Think of the civic virtue evident in the environmental movement, or the women's movement, or the civil-rights movement. Think of the virtue evident in people's being willing to give time to public affairs, to stand for office, and to exercise office punctiliously. Were such virtue not available in a society, then there would be no hope of keeping the law in line with people's attitudes and in ensuring that they remain the masters of the law, not just its subjects. Adam Ferguson[14] was firmly of the view that the rule of law in Britain of the eighteenth century was a great force for republican freedom. But he emphasized the bedrock of virtue on which it must rest if it is to have that effect: "it requires a fabric no less than the whole political constitution of Great Britain, a spirit no less than the refractory and turbulent zeal of this fortunate people, to secure it" (*ibid.*, p. 167).

---

[14] See also my *Republicanism*, ch. 8.

[15] Jean-Fabien Spitz, *La Liberté Politique* (Paris: Presses Universitaires de France, 1995).

A third reason why the rule of law depends on a regime of civic virtue is that there is little hope of ensuring compliance among ordinary people, or among public officials, unless people generally are prepared to remain vigilant in their inspection of one another and in their readiness to rebuke or report offenses. The price of liberty, in the old adage, is eternal vigilance, in particular vigilance in looking at those in power and in challenging, where necessary, their claims and initiatives. A paradox worth noting is that the more people remain ready to rebuke and report in the manner envisaged, the less occasion there may be for their having to do that. To the extent that people know they are all disposed in this way, they each know that others will think badly of them for offending against the law and think well of them for complying with it. To that extent, they will become motivated toward compliance by the desire to avoid shame and win renown. Here is a familiar republican theme: where virtue fails, the saving vice may be a love of glory.

This case for the necessity of virtue in republicanism, it is worth mentioning, ought also to be available to liberals who think of freedom as noninterference, not as nondomination, and who see the institutions designed to support freedom as causally rather than constitutively related to it. But as a matter of fact, the emphasis on virtue does distinguish the republican from the liberal tradition. This is probably because the emphasis in liberalism on minimizing interference, including the interference of the state, has tended to argue against any moralistic interventions by the state as well as against interference proper. The ideal model of government for liberals has often been the government of the invisible hand that is allegedly exercised by the free market. The market takes people's preferences as given—it is not designed to be formative, in Sandel's phrase—and arranges things so that, if people pursue their own individual preference satisfaction, there will still be a collectively satisfactory result: people will sell their services and their products at the competitive price. The holy grail of liberal politics has always been a mode of organizing government which would enable the state to be equally indifferent about the supply of virtue. The goal, to connect up with Sandel's main theme, has been a no-value neutralism.

---

[14] *An Essay on the History of Civil Society* (Edinburgh: Millar and Caddel, 1767; reprinted, New York: Garland, 1971).

III. SANDEL'S REWORKED NARRATIVE

Suppose we now go back to Sandel's narrative and read it under the assumption that the three ideas introduced in the last section are endorsed. What effect is this going to have on that narrative? I want to suggest that the reworking required will have little or no impact on the first three themes in Sandel's story and that, repairing the indeterminacies noted, it will make for a considerable improvement in the fourth and fifth. I shall first consider how it can make for that improvement in the last two themes and then go back to look at how it sustains the first three themes.

(1) Sandel's fourth and fifth themes. The indeterminacies that worried me in Sandel's account of America's lost republicanism bear on the nature of republican ideals, the nature of the civic virtue they require, and the nature of the institutions and programs that those ideals, and that virtue, would mandate. As for the nature of republican ideals, a reworking around the elements I introduced would say, first, that the ultimate ideal is freedom as nondomination for the citizens of the republic, where citizenship is an inclusive category; and, second, that the freedom of citizens is constitutively tied to those participatory and other institutions whereby it is supported: the invulnerability of citizens to arbitrary interference, insofar as that ideal has been achieved, is constituted by their being embedded in such protective and empowering practices. There is no secret about the nature or attraction of republican freedom, under this account, and there is no mystery about how it can be constituted for citizens by the public institutions under which they live. So far it should be manifest, then, that we have attained a clarity that is missing in Sandel's own formulations.

As for the nature of the civic virtue required by republican ideals, that also becomes unmysterious under the proposed reworking of Sandel's narrative. We saw in the last section that, as republican wisdom has it, there are many ways in which the participatory and other institutions required to support republican freedom are going to depend for their existence and effectiveness on a supply of various sorts of virtue among the citizenry. Allegiance to and compliance with the practices in question; readiness to assume a public profile, even a public office, and to try and ensure that the practices continue to answer to the ideas or interests of relevant parts of the society; the vigilance and courage required to ensure that those who disrespect the practices, particularly those in power who do so, are called to account. These are salient examples of the sorts of civic virtue that republican institutions are bound to require—again, no vagueness, and no mystery.

What, finally, of the direction in which republican ideals and republican virtue must be expected to take the state? What of the institutions and programs they would support? Under Sandel's original narrative, we are offered little or no reason to identify any particular direction as presumptively attractive. Sandel himself is clearly of a decentralist, a reformist, and a progressivist inclination: he favors Tocqueville's mediating institutions over Rousseau's single assembly; he thinks that inequality should be reduced; and he believes that a republic would be tolerant of homosexuality and, more generally, would be inclined to uphold individual rights. But his remarks on these matters of policy are few and far between and they often seem to come of independent commitments, not of any firm sense of what republicanism requires.

As with the other indeterminacies, I believe that this all changes utterly—but still, from Sandel's point of view, congenially—under the reworking I propose. When Sandel sets up republicanism against liberalism—that is, against no-value neutralism—then it appears that he espouses a nonneutralism, welcoming into the public forum all the discontented, moralistic voices that are currently marginalized; that is inherent in the prospect of a self-governing community where nothing is taken off the agenda in the name of neutrality. Thus, we are naturally left in some uncertainty about where republican government is likely to lead and in some doubt about Sandel's belief that it will lead in a decentralist, reformist, or progressivist direction. But such indeterminacy is banished once it is clear, as under the proposed reworking of Sandel, that while republicanism rejects no-value neutralism, it does this because of explicitly giving one overall value—freedom as nondomination—a guiding role for law and government, not because of wanting to let loose the dogs of moralistic enthusiasm.

If liberalism represents no-value neutralism, then republicanism represents what we might call a *shared-value neutralism*. The idea long embraced in the republican tradition is that everyone is bound to cherish the prospect of freedom as nondomination for themselves and so, forced to universalize, are bound to embrace the neutral value of freedom as a very important good. But though it is a good that ought to appeal from almost every perspective, freedom as nondomination is something that can be promoted for all only on the basis of collective, political action. So the public institutions of a good society ought to be designed with a view to advancing this value, if necessary in the teeth of opposition from more sectarian or sectional commitments.[17] What I think of as the republicanism of the commonwealthman tradition has been described, for this reason, as an old liberalism that stands in stark opposition to the new: as a shared-value, rather than a no-value, neutralism.[18]

Once we see republicanism as committed to the promotion of the ecumenical value of nondomination, but not independently associated with any particular value position, then it becomes possible, in the light of empirical assumptions, to work systematically at elaborating the direction in which a republican government ought to go. Domination is a threat in our society on two broad fronts: in virtue of the private power or *dominium* that some individuals or organizations or groups exercise over others; and in virtue of the public power or *imperium* assumed by those in government itself. We can investigate where republicanism is likely to lead by looking at the measures of protection, empowerment, redistribution, and recognition that are required for combatting *dominium*, and by reviewing the different controls, the different checks and balances, that are going to be necessary if the *imperium* of government itself is not to represent a dominating presence in our lives. There is a sensible research program here and, though the research clearly remains to be done, still there can be no serious complaint that the policy direction of republicanism lies shrouded in permanent mist.

So much for the way in which we can draw on the elements introduced in the last section and rework Sandel's narrative so that the indeterminacies I identified are repaired. This done, the fourth theme in that narrative assumes a new character: the republicanism to which he would direct us takes a firmer and somewhat different shape. The effect of the change is immediately apparent, as we turn to the fifth theme: the recommendation that we espouse republicanism as an alternative to our current liberalism. There is no longer any need, with the doctrine we now discern, to rest our case for embracing it on the revulsion that we should allegedly feel for the liberal alternative. There is clearly much to be said for the value of freedom as nondomination—much of it continuous with the things that are often said in support of liberalism—and there is room for exploring the advantages that argue for a republican as distinct from a liberal public philosophy.

---

[17] See my *Republicanism*, ch. 3.

[18] Christine Korsgaard, "Commentary on Cohen and Sen," in M.C. Nussbaum and A. Sen, eds., *The Quality of Life* (New York: Oxford, 1993).

(2) *Sandel's first three themes.* What I recommend is a reworking of Sandel's narrative, not, so I claim, a complete rewrite. But if I am to bear out that claim, then I need to show that the changes I have proposed in thinking about republicanism are consistent with the story that he tells in the first three themes identified earlier: in effect, with the story that most of his book is given to elaborating. I turn to that topic in concluding.

His first theme is that American constitutional law has become rights based in a way in which it was not originally so and in a way in which it would not remain so under a republican renaissance. There is nothing in my proposed account of republicanism that undermines the historical story that Sandel tells; for all that that account implies, the story may be fundamentally correct. But would my account of republicanism sustain his sense that the no-value neutralist mode of constitutional thinking represents an infidelity to America's founding republicanism?

I say that it would, for reasons that are many and manifest. Take his remarks on the civic as distinct from the voluntarist approach to religious freedom, to freedom of speech, and to privacy. The fact that religion is an area of presumptive importance in social life, giving people an identity that is significant for them, implies that the state should protect people in this area in a special way, not just by way of honoring personal preference; for any exposure in this area to arbitrary interference on the part of others is going to have a very serious effect on people's freedom as nondomination. Again, the fact that some speech is essential to making one's voice heard in the public forum, and essential to seeing that the public authorities continue to track the interests and ideas of those in one's quarter, means that that speech calls for protection on quite a different basis from other sorts of speech; indeed, other sorts of speech, such as that associated with pornography or even burning the flag, may be thought detrimental to others' freedom as nondomination, or to the infrastructure of that freedom, and may not call for protection at all. Or, finally, the fact that privacy in some matters is an essential protection against certain forms of domination, and privacy in other matters not so essential, means that privacy will not be defended on a blanket, voluntarist basis but in proportion to the way in which it serves the ultimate republican cause.

Take a particular issue like that which is raised by homosexuality, therefore, and consider the way in which the reworked version of Sandel's republicanism would support toleration. The support would not be of the liberal, neutralist kind that he criticizes: the kind that would say it is not the state's job to impose community values in a case like this and so that the neutral state has to stay its hand and respect the privacy of practicing homosexuals (106-07). The republican state would argue for toleration of homosexual practices on the basis of the shared value of freedom as nondomination and would argue with those of more moralistic views that their opposition is inconsistent with espousal of that central ideal. Thus, it would address those views, as Sandel requires, and would stand some chance of winning for homosexuals more than the "thin and fragile toleration" (107) which is all that liberalism can offer.

Take again an issue like that of whether abortion ought to be available to those women who request it. Sandel is forthright in his claim that the neutralist approach is just no good here as a basis for toleration: that in face of the religious values that people so passionately invoke on the other side, the case for toleration cannot be effectively made in a neutralist key; the only effective route will be to fight fire with fire, and to combat value with value (21). Sandel does not say where he thinks republicanism would lead on this issue—in my mind, a worrying silence—but it is worth noting that on the reworked republicanism I would have him endorse, the case for toleration can be defended in the manner he seeks. It is clear that making abortion unavailable would mean that women are dominated by the state—they are subject to a sort of interference that is not forced to track their interests and ideas—and it would clearly expose women to other forms of domination by partners and families and communities. What, however, of the fetus? Here, I must be controversial. I simply record that the most compelling view, as I see things, is that until the fetus is viable, it cannot assume the place of a citizen and cannot make a claim on the protection and empowerment that the republican state would offer to its citizens.

As the reworked republican narrative fits well with the first, major theme in Sandel's book, so it fits well—if anything, better—with the second. Sandel documents the concern of American republicans with industrial employment and the dependency that it introduces for employees. He is forced, because of his emphasis on self-government and the virtue it requires, to argue that this concern stems from a worry that dependent workers will not develop the character necessary in the citizens of a self-governing republic. But this argument always seems strained in the account he offers and is never really substantiated by his many quotations from the speeches and pamphlets of earlier republicans. They complain about dependency as such: about the plight of the "man whose bread and that of his children are dependent on the will of any other man, or who has no interest in his work except to please an employer" (187). It seems artificial to construe the complaint as an expression of concern for the quality of the citizenry and the prospect for self-government.

There is no need to construe the complaint in this way, if we adopt my account of republicanism and take the concern with freedom as nondomination to have been at the heart of republican worries. On this account, the dependency of employee on employer is almost bound to give the employer a certain power of arbitrary interference in the life of the employee—if only the interference implicit in the power of firing without good cause—and so employees are going to assume the profile of unfree, dominated agents. They are going to seem no better than slaves in the extent to which they live at the mercy of employers and in the degree to which they have to bow and scrape to "that haughty, overbearing disposition, that purse-proud insolence" that the Lowell protestors found in their masters (153). They are going to be presented, as in fact they were recurrently described, as "wage slaves."

Not only would the reworked version of Sandel's narrative recast this concern with dependency in a more natural, more persuasive, manner; it would also enable us to sustain Sandel's sense that consumerism betrays republicanism. If we are concerned about freedom as nondomination, then we are going to look to economic arrangements for much more than the satisfaction of whatever preferences consumers display. Conditions of production are going to be important, since, as we have just seen, they can reduce employees to the status of dominated dependents. Issues of poverty and inequality are going to be significant to the extent that they can, and clearly do, make people vulnerable to the evil of domination against which the republican state must set itself.

It is worth noting here, in passing, that the republican state that I envisage would not only have reason to concern itself with dependencies and vulnerabilities that the liberal state typically ignores. It would also be proof against the standard liberal or libertarian argument that when the state does anything about problems of private interference, it itself must interfere in people's lives—at the least it must tax people to fund its activities—and so everything depends on whether the interference perpetrated will be justified by the interference prevented. From the republican point of view, the danger to be eliminated is domination—exposure to arbitrary interference—but the means adopted by the state in the attempt to eliminate such domination need not be themselves dominating: they will not be, if the interference involved can be effectively made out to be nonarbitrary. There will be significant limits on what the republican state should attempt, since it is clear that an excessively powerful state can easily become dominating; and there will be a recognition of the fact that, even if law does not itself dominate, it does reduce the ease or range of choice in which people can enjoy nondomination: nonarbitrary law offends in a secondary way against nondomination[19] But there will not be the same inhibition about state activity which libertarians would impose.

Finally, to the third theme in Sandel's narrative, which is implicit throughout his book but hardly gets any explicit mention. This is the idea that the interest-group pluralism that characterizes contemporary American politics is itself neutralist in having the stronger side win, regardless of the deliberative considerations or values that argue the other way. If freedom is nondomination, and if nondomination is not to be compromised by governmental activity, then there must be constraints in place helping to ensure that such activity will track the interests and ideas of every sector of the society: or, assuming that the sector wishes to belong to the society, at least those interests and ideas that are consistent with their wishing to live under a shared political arrangement with others. But it is surely clear that under interest-group pluralism, there is no hope of achieving this; such an arrangement would represent, not a nonarbitrary mode of decision making, but the rule of naked preference and crude power.[20] If it is to hold out any chance of avoiding domination, the political system must come to represent a rule of reasons, and the democracy come to display a deliberative rather than a bargaining character.[21]

As the reworked republicanism would set us against a purely rights-based law and a purely preference-based economics, so therefore it would set us against the third, less salient of Sandel's demons: a purely interest-based politics. I conclude that the reworking proposed for *Democracy's Discontent* is indeed just that. While it would certainly make for changes in some quite central claims about republican freedom, republican virtue, and republican policy, it would sustain most of the themes that the book puts on parade. The themes might march to a different drum, but they would continue to march in the same direction.

---

[19] See my *Republicanism*, pp. 76-77.

[20] See Sunstein, *Democracy and the Problem of Free Speech*.

[21] Joshua Cohen, "Deliberation and Democratic Legitimacy," in A. Hamlin and P. Pettit, eds., *The Good Polity* (New York: Blackwell, 1989), pp. 17-34.

PHILIP PETTIT

Australian National University
Columbia University

BOOK REVIEWS

Ethics in the Public Domain: Essays in the Morality of Law and Politics. JOSEPH RAZ. New York: Oxford University Press, 1994. 376 p. Cloth $69.00, paper $29.95.

Ethics in the Public Domain is a powerful collection of essays on morality, politics, and law. The essays are divided into two parts: the first contains essays devoted to well-being, and the second includes essays on the relation between law and morality. The essays are intended as illustrations of Joseph Raz's previous writings, both in the sense of explications and in the sense of exemplifications. The illustrations are such an integral part of Raz's thought as a whole that it might be appropriate to give them the medieval name *illuminations*.

Raz's book is outstanding in its high rate of arguments per page. The footnotes squeeze an even higher rate of arguments into small print. This has a price, namely, that Raz's philosophical picture, which I believe is rich and important, is often covered by a thick layer of arguments. Moreover, trying to extract the picture from the collection of arguments incurs the danger described by Ludwig Wittgenstein's lovely metaphor of extracting the raisins from a raisin cake. A raisin cake is a good thing, while a heap of raisins is too sweet and sticky. I can only present a heap of raisins here; the reader who wants the flour as well will find Raz's arguments in the book.

Raz's philosophical picture is of the good life, which is organized around the idea of well-being. Political ethics is concerned with the good life; its aim is the advancement and protection of people's well-being. Raz is not neutral about what the good life is, nor does he expect the state to be neutral in its notions about the good life of its citizens.

Raz is a *perfectionist liberal*. He is a liberal in the sense that he believes freedom plays a central constitutive role in the good life, and his respect for freedom is based on this belief. In his seminal book, *The Morality of Freedom*,[1] the idea of freedom gets the most stress in his depiction of the good life. In the present book, two other emphases are added: the importance for the individual's well-being of belonging to a cultural group, and the role of an active life (as opposed to a passive life) in the good life.

Belonging to a cultural group gives *content* to the individual's freedom of choice. It determines (for good or for bad) the set of alterna-

---

[1] New York: Oxford, 1986.

0022-362X/98/9502/97–104

tives from which the individual may choose. The culture is particularly important for determining the life options available to the individual. The culture plays a double role here: as conferring meaning, and as bestowing value on the available options. Thus, the individual's evaluation of the options is crucially influenced by her belonging to a cultural group or groups. (An individual can belong to more than one cultural group at the same time—for instance, she can be Irish, Catholic, and a member of the working class.)

The second notion emphasized in the present book, that of activity, is also relevant to freedom. Raz is not a philosopher of life but a philosopher of the good life. When he speaks of activity he is not referring to the energy that keeps life going—*élan vital*, or *conatus*, or libido, or the primeval will. What he means are actions based on reasons undertaken to achieve goals that the individual has set for himself. The goals that constitute the good life are not necessarily global ones that cover the entire course of life, or even personal "five-year plans." The main point is that our lives are good ones because of what we do and not because of what we are. For Raz, activity is important in the good life not because it fills us with adrenaline and keeps us on the go; to the contrary, meditation is also a form of activity according to Raz's definition. The most important aspect of activity is the element of the individual's control over his own life. The sort of activity that leads to a good life must be wholehearted. By this he does not mean to rule out activity that we approach ironically or even ambivalently. He only intends to exclude activities that make us hate ourselves. Self-respect, in Raz's view, necessitates the absence of self-hatred and shame about what we do and who we are.

The good life, as Raz sees it, is not necessarily a life free of pain and suffering. Olympic swimmers feel intense pain during training, and Alpine mountain climbers suffer from cold and lack of sleep, but the challenges they face are worthwhile to them, justifying the pain and the cold. Moreover, these are endeavors which are not guaranteed against agonizing failure. The great value of these endeavors lies precisely in the fact that they have such great potential for failure. The good life requires worthwhile goals; and not only does their worth not guarantee them against the possibility of failure, it is internally related to this possibility.

A society is a good one if it fulfills its obligations toward the well-being of the individual. These obligations, according to Raz, boil down to promoting conditions under which individuals have available to them the basic opportunity to strive toward their goals and develop personal relationships in such a way as to make their lives fulfilling.

The good society does not have to guarantee its members success or prevent them from failing. But it is obligated—as are its members as individuals—to guarantee everyone in the society the *possibility* of succeeding by striving to realize worthwhile goals. But even if we have these duties to the well-being of others, in the final analysis their well-being and their success in leading a worthwhile life are up to them: nobody can take over their life and make it prosper for them. This view of Raz's is meant to be not just an instance of "positive thinking," but a conceptual truth. Just as no one can live your life for you, so no one can live the good life for you. Your life is good if *you* live it that way.

I am not sure that this is a conceptual truth; I am not even sure that it is true. There are people who live their lives vicariously. A mother whose two children are the presidents of prestigious universities can live her life through their success, without there being anything in her own life which would give it more meaning than her children's success. I would not deny that she has a good life because of this, even if it turned out that she was not the person who raised the children and so cannot be basking in the light of her past efforts.

Raz's perfectionism, which places the good life at the center of his thinking, is not related to the distinction between the "high" life and the "humble" life. Whether one's life is "high" or "humble" affects the life options available to the individual, but even the "humble" life does not deny the person a choice of worthwhile options. At any rate, Raz does not free us of the necessity to make value judgments related to the good life. Such judgments are just as important for understanding ourselves and others as are character judgments.

What provides depth to Raz's picture is the dialectic he suggests for the relation between the individual and the society—a dialectical relation that undermines the pat dichotomies between personal interests and the moral demands of others, between the rights of the individual and the society, and particularly between egoism and altruism. The bridging concept is that of the common good. Raz opposes Amartya Sen's shrewd observation which seems to advocate separating the concept of rights from the concept of interests and considers it impossible to reduce rights to interests. I have total rights to my old shirt. I would consider myself harmed if someone took it away from me without my permission. This reaction is not due to my interest in the shirt itself, which is negligible, but due to the harm to my rights to the shirt. More important than my right to my old shirt is my right of free speech. There is a considerable gap between the enormous importance we attribute to the right of free speech and the interest of individuals in free speech (except for a few "talking heads" whose interest in speech is intense). Raz does not separate rights from interests; he broadens the concept of interests so that this separation does not occur.

Each of us has an interest in public goods. These are not only the social goods which economists discuss, such as street lighting, which benefit everyone, including free riders who do not pay for them. The social goods of the economists are social goods only in the instrumental sense, in that no one has been able to find an effective way of preventing free riders from using them. The public goods that interest Raz, in contrast, are *essentially* social goods. Living in a democratic society, living in a liberal society, living in a cultured society, and living in a tolerant society are all social goods in which each one of us has a vital interest. I cannot estimate the importance of free speech for myself only according to the degree that I myself need to be able to speak freely, but according to the contribution of this right to the social goods in which I have a clear interest, even if it is not a personal interest in the narrow sense.

The close link between the rights and the interests that we have in public goods bears significantly on the relation between law and politics. That is, the impossibility of separating personal and public interests (which are also public goods) implies that it is impossible, both in practice and in principle, to prevent the courts from becoming involved in political issues when they discuss legal issues bearing on rights. This fact also affects the relation between law and morality, which is the subject of the essays in the second half of the book. The reason is that, in Raz's view, the nature of the law can be understood only against the backdrop of a political society that tells us what are considered public goods.

Belonging to a cultural group is a public good of supreme importance for many people, as such a group is a source of identity and a focus for identification. A cultural group is an encompassing group, one which provides a way of life that includes many important aspects of its members' lives. Being a member of this sort of group is based on belonging rather than achieving. You belong to the Irish if you *are* Irish, and not because of anything you have achieved. This is especially important nowadays, because it serves as a compensation for many people without personal achievements in a society where the market stresses achievement.

But when the culture provides its members with such a vital service, there is a danger that it will be totally guarded against any change that might undermine the group's identity and, even more so, its members' identification with it. Consider the example of a multicultural society. There is a clear danger that such a society may become conservative, if not reactionary, in its attempt to safeguard the identity of the cultural groups in the face of any change. This is not Raz's view. He is not out to protect fossilized cultures that harm the well-being of their members. His notion of multiculturalism concurs with change, including radical change within cultural groups. For Raz, what makes change possible is granting opportunity for exit to those who no longer want to belong to the group. They can thus dare to take the risk of changing their culture from within.

Raz's approach to multiculturalism is not conservative, but, unlike postmodern approaches, it is not fashion oriented either. Postmodernists look at cultures, including the scientific culture, as fashions. Just as there is no external justification for fashions in clothes, there is no such justification for cultural fashions. Philosophy itself is seen as a fashion, requiring self-irony to understand itself as a passing fad. In this view, multiculturalism is a means of pointing out the nature of culture as fashion. One fashion's claims of superiority over another should not be taken seriously, beyond the fact that it is the accepted fashion of a certain number of people here and now.

Raz does have some postmodernist sensitivities in his attitude to the "other," the different. He also places a postmodern sort of stress on the context dependence of political thought and political morality. Moreover, he anchors political morality in existing practices. But Raz is a radical, systematic logocentric thinker who always connects different areas of political morality, placing them under a small number of central concepts. He does not rely on a random collection of intuitions in response to fashionable problems that are considered relevant. At any rate, the reason for defending and encouraging cultures is not based on considerations of skeptical relativism, which claim that one fashion is as good as another and no one has the right to judge them. Raz's reason is the vital role that culture plays in the well-being of the individuals who belong to the group. In order to belong to a cultural group, you do not have to believe that your culture is superior to others, just as you do not have to think that your children are the smartest and the best in order to love and take care of them. Not only is your belonging to a cultural group not based on *your* achievements; you do not choose the culture to which you belong on the basis of *its* achievements but simply on the basis of your belonging to it.

This view is especially important when the cultural group is a nation. Belonging to a nation does not require a belief in its superiority. In a paper that Raz and I wrote together (which appears as chapter 5 in the present collection), we discussed the national right to self-determination. We dealt at length with the question of which groups need to be taken into account as candidates for the right to self-determination. We came to the conclusion that there are quite a number of encompassing groups that should be considered as candidates for this right, with the nation being only one of them. We pointed out the importance of belonging to such groups in the life—and particularly the good life—of their members. But we did not consider the right of such groups to political expression as essential to their living the good life. Such expression might sometimes be instrumental for promoting the prosperity and well-being of cultural groups, but national sovereignty over a particular territory as an expression of self-determination is only a means of expression for a group, and not a necessary condition for the promotion of the group's well-being.

Freedom depends on options. The freedom to be a garage owner or a doctor depends on the rules that constitute the options of being a garage owner or a doctor. Among the rules that constitute options—especially life options of the sort of being a garage owner or a doctor—is the law. But what is the law? In what way do its rules differ from other social rules?

The opening chapter of part II of the book, which is concerned with morality and law, offers an answer to this question. The general tendency is to answer the question of what the law is only as it is seen from the viewpoint of lawyers. This tendency is obviously not coincidental. Most of those who are interested in the question of what the law is are people who were trained as lawyers. What, then, is the law from the viewpoint of lawyers? Raz puts it this way: "The law has to do with those considerations which it is appropriate for courts to rely upon in justifying their decisions" (183).

Raz believes that there is something inherently implausible in the viewpoint of lawyers when it is examined methodologically. The actions of lawyers and courts are parasitic on the broader political and institutional context. Only this context can clarify what the law is, since trials are only one particular way of resolving conflicts in society. This controversy is not one of those tedious philosophical debates between "nothing but"—the view that the law is nothing but the considerations of the courts—and "something more"—the view that the law is part of a broad political system of providing authoritative adjudication for resolving conflicts. The "something more" is a change in the viewpoint and not merely the scope of the issue. It is a change from seeing the law in the context of the courts to seeing it in a political context.

This shift is supposed to shed light on the relation between law and morality, where the courts are seen as a political institution. Raz defends his view that "law consists only of authoritative positivist considerations" (190) by an illuminating analogy from individual to social action. Two stages can be distinguished in individual action: the *deliberative* and the *executive*. In the deliberative stage, the agent considers the advantages and disadvantages of various lines of action. This stage ends when the individual forms an intention to act. If the agent also makes a decision at this time, this means that she is opposed to re-opening the deliberative stage.

This division into stages applies to the actions of organizations as well as individuals. Ethical considerations are likely to be important in the deliberative stage, in connection with what I or we ought to do. The executive stage, in contrast, is conceptually a purely authoritative, positivist stage. From the viewpoint of the courts, laws are authoritative, positivist instructions about what the courts should do, so that in this sense the law is part of the executive stage. The courts, for their part, deliberate about other considerations, including moral ones. Where their deliberations touch upon the law, however, they take it as a set of positivist instructions for execution handed to them by political systems, such as the legislature, the government, or the church.

This picture of the two stages—the deliberative and the executive—is obviously very schematic, but it can still illuminate the way the law obtains its positivist character even though the courts and the legislature must also take moral considerations into account.

The issue of legal authority has been an important part of Raz's thought from the outset, as can be seen from the name of his earlier book, *The Authority of Law*. The present book addresses the additional problem of justifying the duty of obeying the law. Raz, like his teacher, H. L. A. Hart, lives in a world that is disenchanted with the law. Both of them tend to lower the status of the law from ceremonious loftiness to dry prosaicness.

Raz separates the question of who can be taken into consideration as having legal authority from the question of who ought to have such authority. The first question is not decided on the basis of moral considerations: "Nazi rules...are the sort of thing that can be authoritatively binding, whereas statements about volcanoes cannot" (202). I am not sure that this is right—that volcanic eruptions cannot have a normative status. After all, when the commandments of the Torah were given on Mount Sinai, the description of the source of their authority is full of volcanic imagery: "And mount Sinai was altogether on a smoke, because the Lord descended upon it in fire; and the smoke thereof ascended as the smoke of a furnace and the whole mount quaked greatly" (Exodus 19:18). I do not believe that it can be determined a priori which statements will be irrelevant for deciding who is capable of having legitimate authority.

At any rate, it is clear that the accepted origins of the law—legislation, court judgments, and custom—are outstandingly capable of fulfilling the conditions for granting legal authority. This is because they say what subjects are supposed to do, and they can be identified as authorities independently of the considerations they use in their decisions.

AVISHAI MARGALIT

Hebrew University/Jerusalem

THE JOURNAL OF PHILOSOPHY

VOLUME XCV, NO. 3, MARCH 1998

BIOLOGICAL UNIVERSALS AND THE NATURE OF FEAR

How are emotional states such as fear to be defined? The problem arises out of a clash of methodologies used in the study of the emotions. In humans, some action-relevant states such as fear count as emotions, while others such as hunger, on the one hand, and respect for the law, on the other, do not. An examination of the ways in which we make the distinction leads me to the conclusion (section I) that we count as emotions only such states as are typically brought on by cognition and reasoning in humans; consequently, states like hunger, which have no such basis, do not count as emotions. But respect for the law does not count as an emotion either; and so I conclude that in order to be included among the emotions, a state must have a long evolutionary history and be something we share with animals.

These demarcation criteria seem to conflict. If the emotions have a long evolutionary history, they are unlikely to be *essentially* cognitive. When ethologists and psychologists say that animals such as geese and rats experience fear and other emotions, they are not suggesting that these states are the products of the kind of cognitive process that in humans typically precedes fear. So how to define fear? If we define it as cognitive in order to accommodate what seems to be an essential feature of human fear, we exclude fear in geese and rats and include respect for the law. If we let these animals in, we seem to include states like hunger which are essentially different from any emotion.

This is where "biological universals" come in. Evolutionary theorists define biological taxa such as species in ways that permit variation, both synchronic and diachronic. The species *horse*, for example, is understood in a way that permits us to recognize that contemporary horses may differ markedly from ancestral horses and that two

0022-362X/98/9503/105–32

contemporary horses might similarly be very different in their properties, and also to hold that all of these individuals belong to the same species. I discuss how these variation-accommodating universals are to be defined (section II). The aim here is to preserve the well-established idea that terms like 'horse' are general terms with divided reference, while accommodating the antiessentialist program of those who insist that species must be individuals, since they evolve. I attempt to broaden this form of classification so as to accommodate the emotions, showing how we can solve the difficulties of section I, while also accommodating some much-discussed cases of variability among the human emotions (section III). My intention here is to be suggestive rather than definitive—it is my pessimistic conclusion that fear is not definable at the present time.

I. DEMARCATING THE EMOTIONS

(1) Emotions tend to cause a subject to respond to specific kinds of situations in a functionally appropriate way; but not all such states are emotions. I might feel hungry when my blood-sugar level gets low, and this would tend to make me look for food, a functionally appropriate response; but hunger is not an emotion. Respect for the law might tend to determine where I park and where I refrain from parking on different days of the week, but it is not an emotion either. There is a question, then, of how one might distinguish the emotions from other situationally specific states relevant to action.

Consider the following pair of situations.

> *Situation A:* I go up in a hot-air balloon. As the balloon drifts upward to 1,000 meters, I glance downward over the edge of the gondola and look at the tiny objects on the ground far below me. Suddenly, I am overcome by panic: though I know that I am perfectly safe, I feel faint, I sweat, my heart flutters. This is fear-of-heights, or vertigo. When I duck my head back away from the side, the panic abates, and in time I can go about my piloting calmly.
>
> *Situation B:* I go up in a hot-air balloon. As the balloon drifts upward, I resolutely stay away from the edge of the gondola, and look at the instruments instead. Glancing at the altimeter, I notice that I have drifted up to 10,000 meters. I recall that the safety limit of this particular balloon is only 7,000 meters. Suddenly, I am overcome by panic: I feel faint, I sweat, my heart flutters. This is fear—fear brought about by the reasoned cognition that I am too high for safety.

It seems intuitive to think that in situation B, I am experiencing an emotion. As described above, the state conforms, for instance, to

Wayne Davis's[1] specification of "experiential fear": it is a state of "involuntary arousal" resulting directly from *the occurrent propositional belief* that harm might occur to the believer.

But what about situation A? My reaction in situation A seems too automatic to count as such. Unlike the fear in situation B, which is brought about by the belief that I am too high for safety, hyphenated fear-of-heights bypasses belief altogether (and thus fails the Davis test).[2] Given merely the occurrence of the right kind of perceptual state (the sight of a chasm just beyond my feet), the state in situation A can occur on the balcony of a penthouse apartment or on a broad but lofty mountain path, even where I firmly believe that I am in no danger. No doubt, vertigo can be closely linked to fear proper in the absence of a definite belief that I am really safe; without such a belief, vertigo will often develop into fear (in part because it is likely to *engender* the relevant belief). But taken by itself, vertigo will persist (or else gradually fade away) independently of what I believe, whereas fear proper will always be linked to belief. In this respect, hyphenated fear-of-heights resembles hunger, cold distress, and the startle reflex: each of these states persists or can be triggered even in the presence of the clear and distinct cognition that the situation does not warrant it. Most philosophers would agree that none of these counts as an emotion.

---

[1] "A Causal Theory of Experiential Fear," *Canadian Journal of Philosophy*, XVIII, 3 (September 1988): 459-84. Davis's cognitive approach is typical of those to which I allude in this section, if more than usually explicit. The cognitive component of the definition is taken for granted by many philosophers. Kendall L. Walton has, for instance, called it a "principle of common sense" that fear involves belief—"Fearing Fictions," this JOURNAL, LXXV, 1 (January 1978): 5-26, esp. p. 6.

[2] In an article in which I otherwise find much to agree with, John Morreall says: "When some joker sneaks up behind us and pops a balloon or says 'Boo!' we do not have to think of a gunshot or rapist to be scared"—"Fear without Belief," this JOURNAL, XC, 7 (July 1993): 359-66, here p. 361. He takes such an occurrence to be an instance of fear, and uses it to argue that fear can be "objectless." Although I concur with Morreall's general conclusions in this article, I do not agree that the state described above is *fear*. Being startled can lead to fear given the right circumstances (for example, circumstances in which one is encouraged to think of gunshots or rapists), but surely the incident as described, if it occurred at a children's party, is not properly described as "objectless *fear*." The startle reflex is a lot more like vertigo, an automated modular response, and it is not connected to fear proper by cognitive complexes. I persist in this disagreement after having considered the arguments of Jenefer Robinson in an important recent article—"Startle," this JOURNAL, XCII, 2 (February, 1995): 53-74. Robinson says that "several (psychologists) have studied the startle reflex as an example of an emotion" (53), adding that most philosophers would find the suggestion "laughable." But her own descriptions of the opinions of psychologists suggests that they are, at the least, hesitant about classifying startle in this way. I shall say more later about my disagreements with Robinson. They are not relevant to my main line of investigation here.

The reason why we make this distinction between fear proper and hyphenated fear-of-heights is, I suggest, that we think that there must be (possibly weak and defeasible) causal connections (a) between *believing* that some appropriate, action-relevant *triggering situation* (S) obtains and the subsequent onset of the emotion specific to that situation, as well as (b) between believing that S does not obtain and the cessation or diminution of that emotion. The states we have excluded are subject to no direct causal connections with belief, but rather to perceptual states of a specific sort. I might, for instance, experience vertigo simply because I am in a certain sort of perceptual state, whether or not that perceptual state leads to the appropriate belief. On the other hand, I shall not experience vertigo simply because I know that I am too high for safety, as in situation B.

Hunger and cold distress display an even more tenuous connection with cognition; they are triggered by bodily states that are not in themselves perceptual. I experience pangs of hunger simply because I am in a certain hormonal state, even when I know full well that I am not in need of food, and the reasoned belief that I am in need of food will not automatically make me hungry. In each of these cases, it is not of the essence of my state to be connected (however weakly) to reasoned cognition, that is, to the all-things-considered belief that the situation warrants it. This, I contend, is why vertigo and hunger do not count as emotions.

These considerations tend to support a cognitivist (more specifically a *doxastic*) view of the emotions; but cognition is not sufficient. Other demarcation criteria suggest that the emotions must be *primitive* states that we share with animals. I have already mentioned respect for the law as a state that is situation specific and action relevant, but unlikely to count as an emotion. Intellectual curiosity is an even more striking example. When one is intensely curious about something—for instance, something one has just discovered in the library or the laboratory—one may become physically aroused—one's heart rate may become elevated, one's breathing labored, and competing physical demands such as sleepiness or hunger may be blocked. Nevertheless, intellectual curiosity is not usually regarded by philosophers as an emotional state: one does not, for instance, find great European savants declaring that excessive curiosity is a sign of the overwrought and primitive condition of some class of humans, as they have held that excessive emotion is a weakness in women and savages. This suggests that it is actually a part of the way in which we conceive of the emotions that they are states or capacities that we share with animals.

(2) The specific problem I want to tackle here arises directly from the employment of cognitive triggering and of ancient origin as demarcation criteria for emotion in humans. Many scientific studies of behavior stipulate that we share some emotional states with animals; psychologists cite comparative studies of "fear" or "anxiety" in mice and other animals; cognitive ethologists talk of "aggression" in primates, birds, and even insects; primatologists and ornithologists investigate social hierarchies and pair bonding. There is no reason to believe, however, that all these animal states satisfy the cognitive criterion. When psychologists talk of fear in rats, for instance, they are talking about unconditioned responses, such as panic in the presence of signs of predators, or conditioned responses as to a flashing light that signals the onset of an electric shock. There is no reason to think that either of these sorts of response is triggered by anything remotely like *reasoned* cognition or belief. These states are triggered in ways much more like our vertigo or cold distress than like the belief-mediated state in situation B.[3]

This yields something of a paradox:

In humans, fear is cognitive.
Fear is the same thing in humans as it is in rats.
In rats, fear is noncognitive.

Or:

In humans, fear is an emotion.
Fear is the same thing in humans as it is in rats.
In rats, fear (being noncognitive) is not an emotion.

Are these triads inconsistent? Some think that they are; for if fear is the same in rats as in humans, how can it be cognitive in humans, but not in rats? As a result, some philosophers are inclined to reject one or other of the statements they incorporate. Those of an *affectivist* bent would generally be averse to the idea that the human emotions are cognitive, and they would cite the continuity with animal states as evidence in their favor. *Cognitivists*, on the other hand, tend to oppose the idea that animals incapable of ratiocination could be subject to the same states as are humans; regardless of other similarities, they say, animals simply are not capable of fear or anxiety, properly conceived.[4] Both positions are subject to difficulties: affectivists are unable to say why hunger and vertigo fail to be considered emotional states, cognitivists exclude animal states.

---

[3] I am assuming the untenability of basing thought processes, such as those involved in situation B above, on classical conditioning. A rat may freeze when confronted with a stimulus that it has been conditioned to associate with an impending electric shock. This is nothing like the process by which I become afraid in situation B. Morreall makes a similar point with respect to fear in the presence of *unconditioned* stimuli: "A greylag goose that has never seen a hawk, and has never experienced harm, responds with fear when a hawk shaped piece of cardboard is passed overhead, but lacks the experience necessary to form any belief about the object of its fear" (*op. cit.*, p. 363).

The problem I am posing is one even for those who conceptualize the distinction between emotions and other situation-relevant states in a way that differs quite radically from mine. Jenefer Robinson (*op. cit.*) insists that the startle reflex is an emotional response, and includes (more tentatively) hunger, pain, and sexual desire; and she allows that a "*passionate* interest in symbolic logic" might count as "an emotional response" (*op. cit.*, p. 66). So her judgments about demarcation are in these cases the opposite of mine. This is because she conceptualizes emotion as a bodily state which renders some feature of the environment salient and which mediates a behavioral or communicative response to the same. This approach to the emotions might be regarded as functional. Mine, on the other hand, is evolutionary: the principle I follow is that the emotions are grouped together because they possess common characteristics that have enabled them to evolve in parallel ways. Thus, I am interested in (though I do not discuss here) the physiological characteristics that make them, but not hunger, open to doxastic triggering. Despite this difference, Robinson does allow (a) that emotions are typically linked to belief in humans (though she does not think that this defines them); (b) that emotions occur in animals, though they are hardly ever, if at all, triggered by beliefs in animals; and (c) that there are action-relevant states in humans that are too intellectual to be counted as emotional. This combination of beliefs suffices to generate, in broad outline, the problem I wish to tackle here.

---

[4] Jeffrey Moussaieff Masson and Susan McCarthy—***When Elephants Weep: The Emotional Lives of Animals*** (New York: Dell, 1995)—say rightly that "Of all the emotions animals might feel, fear is the one that skeptics most often accept and one of the few that comparative psychology investigates" (p. 46). They quote a number of skeptics concerning the other emotions. A biologist is quoted as saying: "[H]uman love differs quantitatively from animal love...[A] human mother is unique because she has an abstract idea of the Good and that therefore human love, unlike animal love, has its ontogenetic beginnings in a spiritual bond between mother and child" (p. 65). The intention of such utterances is to combat the so-called "anthropomorphism" that describes animals in terms of categories that ought allegedly to be restricted to humans.

II. DEFINING SPECIES

(1) I shall be arguing that fear and the other emotions are "biological universals." I shall now attempt to characterize this category by reflecting on some recent treatments of *species*.[5] Michael T. Ghiselin has suggested that species are individuals; David Hull agrees and concludes that species terms must be proper names.[6] I shall attempt to accommodate the scientific concerns of Ghiselin and Hull, while still maintaining that species are a special category of universal and species terms are common names.

Through most of the history of biological taxonomy, it seemed that the most striking thing about species was that the organisms that belong to them were united by observable similarities they bear to one another, and distinguished from organisms belonging to other species by differences more marked than those which prevail within the group. A lion, for example, seems to be more like other lions in qualitative terms than like any tiger or household cat. If one imagined organisms arrayed in a multidimensional space, the distance between any two being proportional to their *dissimilarity*, the organisms of a single species would cluster close together, it was thought, and there would be a definite separation between the clusters that constitute different species.[7] Species were conceived, therefore, as *similarity-based universals*, identified with the region in this "quality space," arrived at by abstraction from its member organisms, and defined by a formula that descriptively specifies the region.[8]

---

[5] I am most directly indebted to Ernst Mayr's classic work, *Animal Species and Evolution* (Cambridge: Harvard, 1963). For a great deal of help with this section of my paper, I am grateful to my colleagues, Bernard Linsky and Alexander Rueger.

[6] Ghiselin, "A Radical Solution to the Species Problem," *Systematic Zoology*, XXIII (1975): 536-44, and "Species Concepts, Individuality, and Objectivity," *Biology and Philosophy*, II (1987): 127-43; and Hull, "Are Species Really Individuals?" *Systematic Zoology*, XXV (1976): 174-91, and "A Matter of Individuality," *Philosophy of Science*, XLV (1978): 335-60.

[7] Notice that defining the kind *lion* using the similarity of lions one to another is implicitly underwritten by the greater difference between any lion and (for instance) any tiger: for if this did not obtain, the similarity criterion would gather in tigers as well as lions. Nevertheless, similarity has been historically more important in defining kinds than discontinuities. Criticizing this, Mayr argues that such unity is illusory. Differences within species can be very marked: age differences, sexual dimorphisms, and so on, can result in a qualitative spread greater than that found between "sibling species" which display a very marked similarity to one another. Because inanimate kinds such as minerals are generally much more unified than biological kinds, Mayr says: "[T]he most characteristic attribute of a species...is that it is separated by a gap from other[s]. The gap that surrounds a species is the core of the species concept" (*op. cit.*, p. 18). According to Mayr, it is not *overall* similarity or difference that is important, but the fact that for any two species there will always be some parameters that display significant differences.

[8] Darwin is frequently quoted as a proponent of a form of this view: "I look at the term species as one arbitrarily given for the sake of convenience to a set of individuals closely resembling each other...it does not essentially differ from the term variety which is given to less distinct and more fluctuating forms"—*On the Origin of Species* (London: Murray, 1859), p. 52.

The idea that species are similarity-based universals falls immediately afoul of evolutionary biology; for

(a) Species evolve, and change their characteristics over time. Regions of quality space do not change with time.

and

(b) The condition for species membership is at least partly concrete and causal: the reproductive offspring of organisms belong to the same species as they. Since this condition applies independently of any resemblance that the offspring might bear to its ancestors, it shows that the definition of species by a set of resemblances is inadequate.

These facts do not controvert the idea that species are groupings of similar organisms; they are, though the similarities that prevail within a species change with time and evolution. They also do not absolutely prevent us from *defining* species in terms of characteristics shared by the organisms that belong to them; for in order to accommodate evolution, one could make the species definition an indefinitely long disjunction of time-bound conditions. For example, one could say that while at time t an individual would have to satisfy condition F in order to belong to species S, at time t' it is the distinct condition F' that determines membership in the same species. This makes the definition a temporal succession of time-independent conditions tracking evolution. Similarly, to accommodate the conditions of descent, one might build a "closure condition" into a species definition, which states that when a belongs to S, the reproductive descendants of a also belong to S.[9]

These maneuvers preserve the definability of species in terms of descriptive formulas; but they miss an essential point. The qualitative homogeneity and distinctness of species is real, but not fundamental in the way that similarity-based definitions suggest. Rather, they are the effects of underlying evolutionary processes. When we take account of the evolutionary facts that constitute the true unity of species, it turns out that species fall into a special class of universals; permitting and explaining similarities, but defined and understood without reference to them. Because they are so defined, the diachronic variability brought on by evolution, as well as synchronic variation, are accommodated without the need for disjunctive definitions.

---

[9] For more on such maneuvers, and on the confusions that have arisen through a lack of awareness of their availability, see Philip Kitcher, "Ghostly Whispers: Mayr, Ghiselin, and the 'Philosophers' on the Ontological Status of Species," *Biology and Philosophy*, II (1987): 184-92.

(2) Before we consider the implications of evolution, let us consider another point. As we shall see below, species figure in explanatory generalizations and predictions concerning individuals. This threatens the idea that they are similarity based. If 'tiger' *meant* 'large striped cat', it would not be right to say that Sher Khan is striped *because* he is a tiger, no more so than to say that I have children *because* I am a father. So 'tiger' cannot be defined in terms of the features that being a tiger is supposed to explain. Rather, it must be defined by reference to features that *cause* tigers to possess these features. *Explanatory universals* are defined, then, in a more demanding way than similarity-based universals—that is, in terms of shared *explanatory* features.

The distinction between similarity-based and explanatory universals is central to the treatment of "natural kinds" pioneered by Hilary Putnam and Saul Kripke[10] in the early 1970s. As the reader will be aware, Putnam argued that no set of observable features would serve to define the application of the term 'water': for no matter how closely a liquid with a different molecular structure ('*XYZ*', in Putnam's famous example) might resemble water in other respects, it would not be water. Only something with chemical composition H_2O counts as water. This argument presupposes that, in the terms discussed above, *water* is an explanatory universal. We do not construct the universal by abstraction; we do not count things as water because they are colorless, tasteless liquids, or similar to other samples of water in just any way. We include them under the designation if and only if they possess the molecular structure that accounts for these and other water features (even if, as a consequence, we end up admitting some samples of water that are not very similar to most others).

This leads us to a problem: the meaning of 'water' seems to have been fixed long before it was known that its chemical composition is what made it water. How can this have been so? Putnam and Kripke proposed that the *extension* of a general term could be fixed independently of any identification of common properties. They proposed that a "natural kind" such as water is defined by ostending a specimen, and stipulating that *water* comprehends *everything that belongs to the same natural kind as that.* Putnam and Kripke thought that the meaning of 'water' did not change subsequent to this stipulation. To accommodate this supposed constancy of meaning, they posited that the *same-kind* relation would be pretheoretic, so that there would be a continuity between ancient uses of the term 'water' and our own.

---

[10] Putnam, "The Meaning of 'Meaning'," in K. Gunderson, ed. *Minnesota Studies in the Philosophy of Science*, Volume VII (Minneapolis: Minnesota UP, 1975), pp. 131-95; and Kripke, "Naming and Necessity," in D. Davidson and G. Harman, eds., *Semantics of Natural Language* (Dordrecht: Reidel, 1972), pp. 253-355.

It is questionable that this could be right, for two reasons. First, it is doubtful that the governing intuition of the Putnam argument, that *XYZ* is not water, could have been shared, or even formulated, by an ancient like Aristotle, since it was no part of Aristotle's own scientific framework that observational properties must be explanatorily derivative. Secondly, there is not one, but many kinds to which a sample of water belongs: hydrogen compound, hydroxide, liquid, and the like.[11] To define water, therefore, we need a more specific same-kind relation than the one that defines these other kinds. I do not know how this difficulty is to be resolved. I note, however, that, if we use not *same kind*, but *same chemical compound*, the problem would not arise. The latter relation is theory laden, however, and it would not maintain the constancy of meaning for 'water' that Putnam and Kripke wanted. I am going to let the constancy condition go here—I shall not worry about how Aristotle could have defined 'water'. Abandoning this consideration, I shall take advantage of the fact that using a theory-laden *same-kind* relation is an effective way to fix the extension of kind terms at the right level.

As molecular chemistry made it clear that for chemical substances 'same kind' means 'same molecular composition', it became evident that the actual molecular composition of water, H_2O, could be used to define its extension. Thus, the ostensive-relational fixing of the extension of 'water' could be replaced by a definition in terms of an explanatory feature possessed by all samples of water. This is an *intrinsic* feature of these samples, something that a sample of water could possess even if no other such sample existed. Generalizing from this example, Putnam, Kripke, and those following their lead assumed that as science uncovered the nature of the same-kind relation in other areas, *intrinsic* properties would become available to define natural kinds.

The proponents of the Putnam-Kripke approach to natural kinds generally assumed that the same conclusions would apply to biological species such as *tiger*.

First, it is an explanatory universal, rather than similarity based. This seems right: many biologists seem committed to the idea that something is striped *because* it is a tiger, rather than the other way around. "An individual belongs to species *X* not because it has the same species-specific characters as other members of *X*"; says one, "it has these species-specific characters because it is part of the species."

---

[11] Here, I am indebted to conversations with Ali Akhtar Kazmi.

Second, species terms cannot be defined in terms of observational similarities. This conclusion is supplemented by the considerations presented in the last subsection. Since species comprehend variety, and evolve with respect to common features, both observational and explanatory, the species term cannot be defined in terms of the properties it possesses at a given time. Further, an animal's reproductive offspring count, for that reason, as belonging to the same species, and it is always possible that the offspring of members of a species will lack any given species-specific character.

Third, the extension of species terms can be fixed by the ostensive-relational method: that is, by ostending a sample and including everything that bears the same-kind relation to this exemplar. I argue below that this is correct, too, provided that the relation used to fix the extension gets it at the right level: *same species* rather than *same genus* or *class*. To ensure that this is so, we shall need to appeal to modern biological systematics to specify the relation.

Fourth, (though no very confident assertion was ever made about this), it was generally assumed that because species are explanatory universals, the ostensive-relational fixing of their extensions would be replaced by a definition in terms of an intrinsic explanatory feature common to all members of the species. Followers of Kripke and Putnam seemed to suggest that this nonrelational "real essence" would be microstructural, as it is in the case of *water*—perhaps some string of genetic code possessed by all and only tigers. It was thought that this feature would constitute a instancehood condition for *tiger*, as *having the molecular composition* H_2O is the instancehood condition for *water*.[11]

We shall see that this last assumption is mistaken. Although species are explanatory universals defined by ostensive-relational method, they are associated with no nonrelational real essence.

(3) I begin my argument by considering another sort of grouping of similar organisms, also important to evolution: the *deme* or local population. Demes consist of individuals that have access to one another for purposes of mating and reproduction. This access is minimally, if at all, restricted by genetic factors. Thus, a deme is sometimes thought

---

[11] The work most often cited in this connection is Stephen Schwartz, "Natural Kinds," *Behavioral and Brain Sciences*, IV (1981): 301-02.

of as a "gene pool" out of which new individuals are created by recombining genes already present.[13]

The organisms that belong to a deme tend to become and stay qualitatively homogeneous through the operation of a variety of causal factors. Within the deme, the same genes get recombined time and again and this sets a limit to the amount of variety that can be introduced when new individuals are created.[14] Natural selection is a homogenizing force as well. Well-adapted characteristics originate in a small number of individual organisms, and then spread through the population because these individuals are more fit than others. Natural selection will also force a qualitative separation between a deme of one species and demes of other species present in the same location. The reason is that, if two populations are competing for precisely the same resources, then the one able to use those resources better will normally drive the other into extinction. Consequently, competing demes tend to evolve in different directions, and to exploit the environment in different ways; that is, they will tend to occupy different ecological "niches." In this way, demes will tend to be both internally homogeneous as well as qualitatively distinct from other demes present in the same place.

A deme is a group of similar individuals, but it is not a universal. It is not an attribute at all, but is more aptly regarded as a collection of interacting organisms, or even as a complex individual which has other individuals (genes or individual organisms) as constituent parts.[15] *Being a member of a given deme*, however, is a shared attribute of the individual organisms that participate in its reproductive structure. Is it an explanatory universal? Can one say, for example, that membership in a particular deme *explains why* this finch has a beak of a particular shape? Clearly, membership in a deme has no *concurrent* influence on the features of the individuals that belong to it—if an individual were to move in or out of the deme, its qualitative features would not change. The deme does not maintain these features or create them in already formed individuals.

---

[13] Mayr defines a deme as "a group of individuals so situated that any two of them have equal probability of mating with each other and producing offspring, provided...they are sexually mature, of opposite sexes, and equivalent with respect to sexual selection" (*op. cit.*, p. 136). The notion of mating is odd here: it does not make a lot of sense that I could mate with another male if only we were "of opposite sexes." It is the conception of recombination that is crucial. My genes are as likely to get recombined with those of a given member of my deme as with those of any other. Mayr allows that this conception of a "panmictic unit" is an idealization.

[14] The so-called fundamental theorem of natural selection, propounded by R.A. Fisher, says that the rate of evolution in a population is proportionate to the variance of fitness in it. This is a good example of facts about evolution being explained by population-level variables.

[15] Ghiselin and Hull take the position that such collections are individuals. Mayr—"The Ontological Status Species: Scientific Progress and Philosophical Terminology," *Biology and Philosophy*, II (1987): 145-66—argues that "population" is a better characterization, while G. Ledyard Stebbins—"Species Concepts, Semantics and Actual Situations," *ibid.*, pp. 198-203—suggests "systems." This particular controversy is not of interest to me in the present context.

Nevertheless, membership in the deme does play another sort of explanatory role. As the influences mentioned above work on a population, the frequencies of genes it contains changes. Such attributes of the deme are among the causal factors that shape the process by which a deme gets homogenized and the member organisms get their attributes. In general, the causal relations that obtain within a deme and those which work on it, and its prior gene frequencies, will be parts of the explanation of evolution. In this sense, membership in the deme explains why its individual members have many of the properties they do. Take an extreme case. A finch has a beak of a certain shape because, as a result of past selection, there are no other genes available in the deme in which it was created. Mutation or immigration aside, there is no other possibility for the individuals produced by the deme. Thus, the finch has this sort of beak because it was produced by this deme. In other cases, the probability may be considerably more distant from certainty, but gene frequencies support an explanation nevertheless. This man has sickle-cell anemia because his ancestors belonged to a deme in which the sickle-cell allele provided some protection against malaria. This tiger is an albino, because breeders ensured that the gene pool that produced it contained as many albino genes as they could manage.

Membership in the deme is an expanatory universal, then, not because it denotes some common nonrelational feature of its individual members, but because the composition of a population, the forces that impinge on it, and the ways in which it produces new individuals predict and explain the types of individual it creates.[16]

(4) A species, too, is a gene pool, though more extended than a deme. It, too, is a composite of individuals that produces new individuals and it, too, is one of the factors responsible for the evolution of property distributions among the ensembles of individual organisms it creates.

A species can be thought of as a collection of demes spread out through space and time, and connected by gene flow. Individuals migrate from one synchronic deme to another, carrying their genes with them, and there is gene flow from temporally earlier to later demes by inheritance. A species is the extended entity within which such gene flow occurs. A species is, moreover, guarded from gene flow from other species by so-called "isolating mechanisms," genetic factors that restrict cross-species mating.[17] A species is thus a self-contained unit. Take an individual organism and include all the others—past, present, and future—with whose genes its genes could recombine by a natural process of sexual mating, and you have a class that constitutes the membership of the species to which the original individual belongs.[18] Thus conceived, a species is, to combine the characterizations of Theodosius Dobzhansky and Ernst Mayr, "the largest and most inclusive...reproductive community of...actually or potentially interbreeding natural populations."[19]

---

[16] We are not obliged here to insist that there are no features shared by all members of a deme (I was assuming above that the beak was an instance of just such a common feature), or even that there are no shared microstructural features. The point is that membership in a deme can be used to explain individual characteristics without appealing to common intrinsic features.

The confusing thing about species is that, notwithstanding the above, they are used in biology as archetypical universals. In fact, their role in biological taxonomy is strongly analogous to that of structural classifications in chemistry. It is not possible to discuss the point in any detail here, but I summarize. First, in most areas of biological science, species categorizations are the primary vehicle for explanatory generalizations and predictions concerning individual organisms.[20] Further, species are unlike demes in that an individual member cannot join or leave a species as it can a deme—since the species is inclusive of all demes within the reproductive, ecological, and ethological barriers that isolate it from outbreeding, movement from one deme to another still leaves the individual within the species. This permanence of species membership supports the explanatory and predictive value of species taxa. Moreover, the logic and grammar of species terms seems to dictate that the relationship of an organism to its species be treated as that of instance-hood. We say "This *is a* tiger,"[21] and if it is also true that "the tiger is a cat," we are entitled to conclude that the thing ostended is a cat. Species are thus associated with the logicolinguistic practices of "*is-a*" predications, which again seems to indicate that they are treated within biology as important categorizations of individuals, categorizations to which explanatory properties are attached. Lastly, species terms have divided reference: they take the plural and are quantified. It makes sense to say: "There are ten tigers in the zoo."

---

[17] These isolating mechanisms evolve as species and their demes come to occupy characteristic niches: the qualitative distinctness that this demands results eventually in reproductive isolation, and thus in the creation of a group that maintains its qualitative distinctness from other such groups and evolves independently. A new species has been created when morphological change within a subgroup of a preexisting species minimizes gene flow in or out, so that the subgroup becomes isolated from genes of the type present in the ancestor group.

[18] Kitcher gives a more complicated definition along the same lines (pp. 187–88). As he notes, the reproductive compatibility criterion is vague.

[19] Mayr quotes Dobzhansky (p. 19). The definition does not apply to asexually reproducing species. The following discussion is meant to apply only to Mayr's "biological species."

[20] Hull and Ghiselin would disagree on the grounds that there are no laws of nature of the form "All *S*s are *F*," where '*S*' stands for a species name. As Alexander Rosenberg—"Why Does the Nature of Species Matter? Comments on Ghiselin and Mayr," *Biology and Philosophy*, II (1987): 192-97—explains: "[T]he generalizations about particular species on which taxonomic decisions rest are full of exceptions, and there is no background theory that will enable us to eventually eliminate, reduce or even explain these exceptions" (p. 195). This is most likely correct, but to my mind it is beside the point. Certainly, generalizations of the form 'Most tigers are striped' support counterfactuals such as 'If this had been a tiger, it *probably* would have been striped', and explanations like 'This is striped because it is a tiger'. Why do probabalistic generalizations not count as "laws of nature?" Why do such generalizations even have to be "laws of nature" in order for the terms that occur within them to be counted as explanatory universals?

It seems, then, that while species are populations just as demes are, they need to be treated analogously with universals with respect to their taxonomic-explanatory and semantic roles. Our treatment of demes suggests how to accommodate this duality. In the case of demes, we distinguished between the deme itself, a causally structured collection of individuals, and *membership in the deme*, a universal. Here, we need to make a distinction between *species populations* and *species universals*. Species populations are composite individuals, the extended reproductive communities of Mayr's definition. The species universal corresponds to *membership in the species population*.[22] The linguistic practices of biologists make the distinction explicitly in the case of demes; with regard to species, they do not.

(5) I return now to the third and fourth characteristics of species universals enumerated in subsection (2) above. The extension of a species term can be fixed by the ostensive-relational method. But the

---

[21] Ghiselin disagrees: "Parts are not instances—we do not say 'This thumb is a Michael Ghiselin'...to some people it is not obvious that it is equally wrong to say 'Trigger is an *Equus caballus*'"—"Species Concepts" (p. 128). I suppose that he is playing on the infelicity of using an article with the Latin species term without interposing 'member of'. Does he think it wrong to say "Trigger is a domestic horse," or "There are thousands of domestic horses on Earth?"

[22] Here, I do not mean to close the door on Kitcher's suggestion that we should be "pluralistic" with regard to species concepts—"Species," *Philosophy of Science*, LI (1984): 308-33. See also the comments of Elliott Sober in "Sets, Species, and Evolution: Comments on Philip Kitcher's 'Species'," *Philosophy of Science*, LI (1984): 334-41. Kitcher proposes that some species classifications might correspond to similarity-based universals (as I call them) or to real essences. As I stipulated in footnote 19 above, my proposal concerning universals is meant to apply only to "biological species," that is, the extended reproductive families discussed above.

pretheoretical "same-kind" definition does not distinguish between species and the other taxa to which the organism belongs—genera, classes, and the like.[23] As we saw earlier, the most sophisticated theoretical investigation of species reveals that organisms are conspecific if they belong to same extended reproductive community. Thus, we may define a species population Σ as follows:

(a) This (ostended) organism belongs to Σ.
(b) If x belongs to Σ, and y belongs to the same extended reproductive community as x, then y belongs to Σ.
(c) Nothing else belongs to Σ.

The species term S is then defined:

$S(x) =_{df} x$ belongs to Σ[24]

The species term thus applies to any individual that stands in a certain relation (*belonging to the same extended reproductive community*) as the originally ostended individual; this relation is given to us by theory. For two reasons, it cannot be redefined in terms of an intrinsic feature shared by every species member: first, though the members of a species may in fact have certain intrinsic features in common, they need not and in the future may not; second, membership in the species is prior from an explanatory point of view.

The status of species as universals and of species terms as common nouns is saved by defining them in terms of membership in a population. Membership in a population is a relational property: an individual possesses this property as a consequence of the relations it bears to other individuals. I have argued that membership in a population can figure in a form of explanation concerning the properties of the members of that population, even though membership corresponds to no common intrinsic feature. Because this is so, explanation by appeal to species membership does not demand the replacement of the relational characterization of species population by one in terms of an intrinsic real essence. This is important. Variability in terms of intrinsic features is a precondition of evolution. Even though populations display common features, any such feature may always be modified in a new organism, and subjected to selective pressures. So, the only way in which the theory of evolution can be reconciled with the traditional conception of species as explanatory universals is if there is a way in which membership in a relationally defined population can be explanatory.

---

[23] On the nature of higher taxa and their definability, see particularly Kevin de Queiroz and Jacques Gauthier, "Phylogeny as a Central Principle in Taxonomy: Phylogenetic Definitions of Taxon Names," *Systematic Zoology*, XXXIX (1990): 307-22; and de Queiroz, "Phylogenetic Definitions and Taxonomic Philosophy," *Biology and Philosophy*, VII (1992): 295-313.

[24] Hull is scornful of designating the species property in the way I have. "One way of salvaging essentialism is to claim that being a member of a class is the essence of the class; e.g.,...being a horse is the essence of horse.... If there was ever reason to suppose that essentialism was no longer a viable metaphysical position, this means of salvaging it is it"—"*Are Species Really Individuals,*" see p. 176. My move is not so vacuous: (a) because there is a distinction between the species population and the species universal (thus there is a difference between defining the universal in terms of the population, and defining the species in terms of itself); (b) because the defining characteristic of belonging to the population is relational, whereas most essentialisms require essences to be intrinsic features; and (c) because it is the status of the population in evolutionary theory that gives us the route to a definition like the one offered. As argued in the text, the extension of chemical terms have no such status in chemistry.

There is, then, a crucial difference between biological systematics and chemistry which makes the transition to a nonrelational real essence unnecessary in the former. The difference is that, in biology, the extension of species terms corresponds to an entity, the population, that enters into explanatory causal relations. In chemistry, the extensions of terms like 'water' have no such ontological status.

I argue in the next section that universals like *fear* possess the same characteristics as those outlined above. Obviously, there are differences: the instances of fear do not, taken together, form a population of interbreeding organisms. Nevertheless, emotions are an inheritable characteristic in respect of which organisms differ, and as a consequence the emotions evolve. They, too, are definable only by the ostensive-relational method. Like species, they have no nonrelational real essence capable of replacing the relation by which their extension is defined. In short, they are biological universals.

III. FEAR AS A BIOLOGICAL UNIVERSAL

(1) The problem that we are trying to solve has variation at its heart. Consider our "paradoxes" above. Fear is something that, because of evolution, takes different forms in different species (indeed in different individuals, and even different occurrences). The first paradox reflects this variability. The doxastic triggering of fear is a qualitative characteristic. However typical this characteristic might be with respect to occurrences of fear in humans, it would be a mistake to build it into the definition in such a way as to exclude all nondoxastically triggered states. Rats are different from humans in this respect.

Our second paradox, too, is the result of biological universals crosscutting against qualitative specifications. Above, we arrived at a qualitative characterization of human emotions, namely, that they are triggered by cognitive states. At the same time, we wished to maintain that in some other animals the same states—fear and the like—could be noncognitive. How are we to accommodate both positions? In one of two ways. We might want to say that though fear is an emotion in humans, it is not truly so in other animals. Such a position would allow us to retain a cognitivist characterization of *emotion*, while allowing that each of the types of occurrence so designated occurs in forms that do not satisfy this characterization. Alternatively, we might hold that the term 'emotion' applies to those dispositions which have a certain characteristic as they occur in humans. Fear in rats is an emotion not because it is different from such nonemotions as fear-of-heights or hunger *tout court*, but because it is the same state as one that is different from these states *in us*. It may very well be the case that the states so designated have a common characteristic (for example, that they are routed through the frontal cortex of the brain), but this feature is interesting only because they permit the doxastic triggering that interests us with regard to human emotion. By the same token, fear-of-heights would not be counted an emotion even in rats, not because its intrinsic character is all that different from that of other avoidance modules, but because of its subsequent history—that is, because it did not become sufficiently connected with cognition and learning.

The paradoxes are the result of evolutionary change. They are no more difficult to accommodate than individual change in a theory of personal identity.

(2) The variable character of fear across species shows us that cognitive characterizations, such as Davis's, are inadequate. The young greylag goose's fear of a hawk-shaped piece of cardboard is not a cognitive state; nor is the chimpanzee's fear of snakes and chimpanzee corpses.[25] Davis's definition fails even within the human species. The child's fear of the dark or of loud noises is not cognitive; there need be no "apprehension of hurt"[26] in it. Even among adult humans, the definition fails. Adults as well as children can be afraid of being alone, of the dark, (like chimpanzees) of snakes, and of dead bodies. They can be neurotically afraid, or even incapacitatingly terrified, of sex or of members of the other sex. Such fears are not mediated by cognition; they may indeed be accompanied by the firm belief that there is no impending hurt in the situation. Nevertheless, they are involved in cognitive complexes characteristic of

---

[25] D.O. Hebb, "Emotion in Man and Animal: An Analysis of the Intuitive Processes of Recognition" and "On the Nature of Fear," *Psychological Review*, LIII (1946): 88-106 and 259-76.

[26] Robert Kraut quotes Collingwood as citing "a woman's fear of a mouse or a boy's fear of the dark" to controvert Thomas Hobbes's view that the "the opinion of Hurt from the object" is an essential part of fear—"Feelings in Context," this JOURNAL, LXXXIII, 11 (November 1986): 642-52, see p. 643.

fear as we know it: as is shown, for instance, by the fact that the fear of any object can be heightened (or even created) by associating it with the dark or with the presence of a corpse. It seems, then, that a number of the modules that are found unconnected in some animals are found bundled together in humans, and triggered by cognitive processing and learning. In a similar vein, the emotions we (humans) experience in response to works of fiction are not doxastic; they occur in the full knowledge that the cognitive attitudes that trigger them are false. Pace Kendall Walton (*op. cit.*),[17] these fearful reactions can be real.

What this variability shows is that it is a mistake to attempt a definition of emotions in terms of shared intrinsic characteristics. It is, no doubt, an extremely valuable generalization concerning the human emotions that they are often, or even typically, doxastically triggered—indeed, I have alleged that it is the human capacity for such triggering that demarcates the emotions from other action-relevant states. But this should not lead us to make cognitive conditions essential to their very occurrence by building such conditions into their definitions. Such attempts are defeated by variation within or across species, and by the explanatory primacy of the collections of things within which natural selection operates.

It has been suggested that a more plausible alternative to a cognitivist definition is a functionalist one. Robert Kraut, for instance, thinks that emotions are "feelings in context," and considers a "functionalist strategy," which treats of the feelings as states which "by virtue of their systematic interconnections and functional organizations, realize the functional roles definitive of the emotions" (*op. cit.*, p. 643).[*] Following this strategy, fear might be defined as a feeling that enables one to respond appropriately to a dangerous situation. This definition is not explicitly cognitive, and so it is inclusive of animal states. Allegedly, it also enables us to accommodate the "nonstandard" emotions such as those we experience in response to fictions. Such states admittedly do not enable us to respond to dangerous situations, but it does not follow that they do not have this function, since not everything is able to achieve its function in every situation. Kraut says: "What renders *F* a feeling of [fear]...is the presence of certain features...in the situations with which *F* covaries under optimal conditions," but he stipulates that under suboptimal conditions, a feeling might occupy the emotional role without these covariant conditions being satisfied (*op. cit.*, p. 649). Optimally, a feeling of fear will enable me to respond appropriately to a dangerous situation, but when I experience fear in response to a movie, or when I am left alone in a funeral parlor at night, it is because conditions are suboptimal.

---

[17] To be fair to Walton, the bulk of his discussion is concerned with showing that in such cases we are not afraid *of the fictional object*. Walton's proponent, Charles, is not afraid of the green slime he sees in a movie, because he knows that it does not exist. So far, I sympathize. But one might ask: If he is not afraid of the green slime, is he really afraid?—Is there anything else to be afraid of? Walton thinks not, and makes his famous move: "The fact that Charles is *quasi*-afraid as a result of realizing that make-believingly the slime threatens him generates the truth that make-believingly he is afraid of the slime" (*op. cit.*, p. 14). Quasi-fear outside the scope of the make-believe operator, fear inside. But I do not accept the 'quasi'. It is make-believe that Charles is afraid *of the slime*. He is, however, afraid period. Perhaps *fear* is an example that gives the Walton example a degree of plausibility: as with being startled by a popping balloon (footnote 2, above), Charles is *scared* or *afraid*, but is he *really* fearful? But this a point irrelevant to Walton; and there are examples less favorable to his analysis. Can we not be truly envious when we read of the enormous wealth of the (fictional) families portrayed by Evelyn Waugh? Are we not truly indignant when we read of the murderous ways of upper-caste Indian landowners in Vikram Seth's *A Suitable Boy* or Rohinton Mistry's *A Fine Balance*? It may be that in these cases, we are led to emotional involvement with real situations, by reflection on the situations fictionally portrayed, and their resemblance to real life. Nevertheless, our make-believe involvement with the fictional case does engage our real emotions: I do not find it plausible to say that the *only* emotion evoked by our reading of Seth and Mistry arises out of our engagement with the *reality* of caste injustice: someone could be indignant even if they had never heard of the caste system.

Does the functionalist strategy really solve the difficulty? How is a suboptimal occurrence of fear to be identified as such, if it lacks central properties of normal fear? Kraut seems to rely on something like Robert Cummins's[*] identification of function with causal role; but since fear of fictions does not occupy the same causal role as "normal" fear, the applicability of Cummins's analysis here is moot. It might be thought more plausible, therefore, to use an analysis of function that tackles nonperformance head-on. According to Ruth Millikan's[*] analysis, an item has the "proper function" of enabling an animal to respond appropriately to dangerous situations if (i) it is a "reproduction" of some prior item that enabled an animal to respond appropriately to dangerous situations, and (ii) we now possess the item because that prior item did so. So understood, a particular occurrence of fear may in fact possess the *proper function* of responding to dangerous situations even though, in fact, it does not manage to do this. It has this proper function because it is a reproduction of an item that...and so on. The difficulty is: What makes an item like fear-in-response-to-fictions a "reproduction" of the prior item? After all, it fails to perform the very service by which we identify that prior item.

---

[*] Kraut abandons the functionalist strategy in the end.

[*] "Functional Analysis," this JOURNAL, LXXII, 20 (November 20, 1975): 741-64.

[*] "In Defence of Proper Functions," *Philosophy of Science*, LVI (1989): 288-309.

The plausibility of analyses like Millikan's (as a basis for functional definitions) rests on their application to items like hearts and eyes. These items are morphologically, and even histologically, distinct from surrounding body parts, and occur in a relatively predictable place relative to other parts of the body. Thus, even a defective heart or eye is easily identified as such. Such cases make it easy to think that we are given a class by perception, the problem being to state what its essential defining characteristics are. (Note, however, that when we do identify a defective heart as such, we do so not by the function but by morphological or other such criteria.) But not all body parts are so easily identified either across or within species. Something like a skeletal muscle is compositionally similar to other muscles, and can only be identified by its attachment to particular bony structures, that is, relationally. Other anatomical elements—for example, the carotid artery—have no definite boundaries, and are identified in an even fuzzier way.[31] A glance at any textbook of comparative anatomy shows how theory-laden identifications of parts are when applied across species. Speaking of pterosaurs, one such text says:

> ...the fourth finger [was] tremendously elongated to form the principal support for a wing membrane.... The fingers in front of the fourth digit were reduced to small hooks.[32]

Thus, similarity is not necessary for identification, and it is not sufficient.

> Animals without jaws...can develop movable plates around the mouth opening that serve after a fashion as weak "jaws"...the appearance of jaws in vertebrates was brought about by a transformation of anatomical elements [namely, gill arches] that had originally performed a function quite different (*ibid.*, p. 30).

Distorted fingers that do not function as such, structures that perform the function of jaws but are not jaws, gill arches that have turned into jaws.... It is clear that the principles used to make these identifications of body parts are neither trivial nor easy to come by. The "same" part differs from species to species in both form *and*

---

[31] For a taxonomy of such cases, see Bruce A. Young, "On the Necessity of an Archetypal Concept in Morphology: With Special Reference to the Concepts of 'Structure' and 'Homology'," *Biology and Philosophy*, VIII (1993): 225-48.

[32] E.H. Colbert and M. Morales, *Evolution of the Vertebrates* (New York: Wiley-Liss, 1991, 4th edition), p. 177.

function. Positional relations, similarities, compositional discontinuities, and ontogenetic criteria have to be used to establish the sameness of part across organisms and across species. Thus, it is not necessarily clear when we find, in an organism of our own or another species, a vessel of a very different shape and function in the place of the carotid artery, whether we are to say that that part has changed its form and function or is simply absent.

When it comes to identifying a fleeting mental occurrence as an instance of *fear* when it is not performing its supposed function, we are in uncharted territory. The analyses of function summarized above are useful enough in attributing functions once the class of fear occurrences has been identified; they are of no use in identifying the class. In responding to dangerous situations, one individual may freeze, another may attack, yet another may warn or get together with members of the same group. Some individuals will be afraid of the dark, some of fictions, some of corpses, some of sex. Is this a case of different emotions, or of the same directed to different objects, leading to different actions? The analyses of function offered by Cummins and Millikan do not offer us an answer. Because they do not, it seems dubious that a *functional* definition of fear is possible, and even more dubious that such a definition will help us identify *optimal* or *normal* instances of fear.

Why is it thought plausible that there are normal as well as subnormal instances of fear (or quasi-fear, as Walton calls it)? Because it is thought that there is something like a real essence of fear, a characteristic that explains its typical manifestations. As Elliott Sober[33] has shown, one maintains this hypothesis in the face of variety by positing interfering "forces" that on occasion prevent or frustrate the working of the essence. This is what the subnormality assumption attempts: it claims that fear sometimes manifests itself in peculiar ways because nonstandard conditions intervene to interfere with its real essence. The biological universal model gives us an idea of how fear can be an explanatory universal without making such an assumption.

This mental event is triggered by the "apprehension of hurt," *this other* event is not. We may wish to explain the characteristics of both by reference to the evolutionary pressures operating on some class of occurrences to which both belong, including (a) the fact that events like the first tend to keep us out of hurt-causing situation, (b) selection favoring this tendency, and (c) various similarities and connections between the two kinds of events that might tend to preserve the second despite its failure to display the selected property. It would be a mistake to think that this kind of explanation, functional though it might be, consists in the attribution to the second event a lesser degree of some capacity that the former possesses in full. Like explanation in terms of biological universals, it reaches down to the individual only through some class to which it belongs, and the variation present within this class. But in order to do this, the occurrences must first be identifed as belonging to the class. By the very nature of the case, functional specification is not going to help us do this.

---

[33] "Evolution, Population Thinking, and Essentialism," *Philosophy of Science*, XLVII (1980): 350-83.

(3) The variability of body parts and of the emotions discussed above suggests that, like species, they are biological universals. As we have seen, the Putnam-Kripke method of ostensive-relational definition helps us fix the extension of universals in cases where characterizations in terms of common properties do not serve. In the case of fear, one would need to ostend (successfully) an occurrence of fear, and stipulate that everything else of the same kind is also fear.

It is here that the so-called "optimal" or "normal" cases are useful. We are interested in the kind of fear described in situation B above: cognitively triggered, leading to classic symptoms, and occurrent in a human. Fear is then the class of occurrences defined by some *same-kind* relation with respect to the exemplar. Provided it is properly specified, this relation will gather in some of the other cases that we have discussed—fear in rats because though not cognitive, it is of the same kind as the exemplar. If we are right in thinking that fear is a biological universal—that is, that it defines a class on which selection has worked, and continues to do so—this relational characterization of fear will not be replaced by a nonrelational characterization, a specification of a characteristic possessed by all and only occurrences of fear. Fear will be understood as constituted by membership in the class characterized as above.

I argued above that a pretheoretical characterization of the same-kind relation would not be sufficient to fix the extension of a term if it did not pick out a kind at the right level. There are many kinds to which an occurrence of fear belongs; we are interested in only that *same-kind* relation that binds the exemplar together with other occurrences of *fear*, not some broader or crosscutting classification. We need a more demanding characterization of the requisite *same-kind* relation if the definition is to do its job. It is the business of empirical science to investigate both the same-kind relation and the class of events defined by it. But it does not seem that empirical science has succeeded in giving us a good way in which to understand homology. The situation is not unlike that which obtained sixty years ago with respect to species. Since that time, the notion of a *biological species* has become much clearer. The same could happen for homology. At present, however, it appears that the same-kind relation is not well understood as it applies to items of this sort—the difficulties alluded to above have not been resolved to the satisfaction of all. The *same-kind* relation is thus a placeholder awaiting further clarification. Until an adequate theoretical understanding of the same-kind relation is achieved, as it applies to body parts and "mental" events, no adequate definition of these things will be possible, even within the framework of ostensive-relational defintions.

David Lewis[34] has argued that mental kinds should be defined functionally, but that in different "populations," they might coincide with different morphological kinds in a way that admits of variation both across and within populations. As test cases, he imagines a "Martian" whose pain is functionally similar to that in humans but physiologically very different, and a human "madman" whose physiologically standard form of pain is functionally aberrant. According to Lewis, the case of the Martian shows that different states can occupy the functionally defined role of pain in different "populations." The case of the "madman" is different: "the state that occupies the role (of pain) for the population (that is, humans) does not occupy it for him" (*ibid.*, p. 127). My argument tells against the functionalist definition of pain, and the considerations of the next subsection tend to tell against the idea that the Martian suffers *pain*. I reject the notion of defining pain by its causal role, and the practice of using this definition to identify the states that in each population occupy that role. But the idea that mental kinds like pain or fear may coincide with morphological kinds within particular biological taxa is one that I can accept, given the proviso that they will do so in a way that admits of variation. Provided that one allows for change, that is, for the possibility that fear might come in the future to coincide with a different morphological kind, the approach that I am taking does support the at-least approximate co-extensiveness of biological universals with some morphological characters. But (much as Lewis intends for his functional definitions) the ostensive-relational definition will be explanatorily prior to these shifting and uncertain liasons with morphological kinds.

(4) Although we do not know exactly how fear will be defined, we do know that the definition will respect descent; that is, some states

---

[34] "Mad Pain and Martian Pain," in *Philosophical Papers*, Volume 1 (New York: Oxford, 1983), pp. 122-32.

will be included because they are phylogenetically connected with the human exemplars, even though they lack some of the characteristics that seem salient in these cases. In such cases, evolutionary taxonomy identifies elements as "the same" across species when they are homologous, that is, when they have a common evolutionary origin.[35] In the above mentioned cases of fingers and jaws, the homology consideration trumps form and function. That a particular structure has a very different shape in pterosaurs does not count against its being recognized as the "fourth digit," that something performs the function of jaws in some *Agnatha* does not make it a jaw. What counts is common origin: the elongated digit has an "ancestor" in common with the corresponding structure in other vertebrates, and the jaw-like structure has a function but no ancestor in common with our jaws. (How "ancestors" are identified in the case of bodily parts is, as we said, a contested question.) It seems possible that the emotions, too, will be regarded as the same or different across species on the basis of homology. This possibility is rich in consequences for the philosophy of mind, some more intuitive than others, and I conclude this essay by gesturing briefly toward some of these.

Emotions are mechanisms that cause an animal to respond to specific situations in specific ways. But it is possible that, just as some creatures use nonjaws to perform some jaw functions, similar responses to a given situation could be mediated by different mechanisms. It is thus not inconceivable that, even if an animal belonging to a distantly related taxon attempts to avoid danger, it might do so by the operation of something other than fear. Within functionalism, it is hard to make sense of this claim: functionalists identify elements as the same when they have the same functional role within a system.

There is an interesting passage in *When Elephants Weep* in which Jeffrey Moussaieff Masson and Susan McCarthy describe a dying mother spider enfolding her young to herself when the entire brood was immersed in alcohol.

---

[35] Philosophers are sometimes inclined to favor a more restrictive criterion than this. In an otherwise interesting article, Crawford L. Elder, for example, makes the claim that kinds like *lung* cannot have unequivocal application *across species*—"A Different Kind of Natural Kind," *Australasian Journal Of Philosophy*, LXXIII, 4 (December 1995): 516-31, here p. 526. I suppose that this would rob me of one of the main motivations of my argument, namely, that fear occurs in many species. But I cannot think what would motivate such a restriction; the proper provenance for a kind based on Millikan's reproductions—Elder's "different kind of natural kind"—is common origin, not sameness of species.

> Can a spider love her babies? Was it a mere reflex that caused the...spider to reach for her young?... It is so hard to have insight into a spider's mind that it is almost impossible to guess.... Yet spiders have evolved to produce complex venoms and digestive fluids.... Building a spider's web is an extremely complicated behavior. One can argue that a spider is not really a simple organism and that the development of maternal love might well be a shorter evolutionary step than web building (*op. cit.*, p. 68).

Masson and McCarthy think that the question is one of evolutionary development: Is the spider *advanced enough* for us to say that it feels love? But there is another relevant parameter which they ignore, and it might be important to the way we answer their questions. Is the evolutionary development along the same path or along parallel paths. Do the two instances of maternal instinct have a common origin, or did they originate independently?

Masson and McCarthy continue:

> What if it was discovered that when a mother wolf spider sees young spiders, her body is flooded with a hormone whose presence is associated with feelings of love in higher animals? Would that be evidence that the spider loves her young? What if it was a hormone peculiar to spiders? Would that mean it wasn't love (*op. cit.*, p. 69)?

The focus here concerns moral questions (broadly conceived). So let us stipulate right away: as far as certain moral evaluations and moral consequences of maternal behavior are concerned, the identity of the hormone that floods the spider's body makes no difference. From the point of view of biological explanation, however, it might make a difference. Because parental investment is an important reproductive strategy, it has evolved independently several times. If spider maternal behavior evolved independently of human, that is, if there is no common ancestor of both forms of behavior, then it may well make sense to say that though the spider's state has the same *moral* significance as love, it is something different.

What sort of consideration would motivate us to make such a distinction? Look again at the hormone question. Functionally comparable items that are not homologous will share functionally relevant features. The dying spider reached out her legs and folded her dying babies to herself in an attempt to protect or comfort. A human mother might do something very similar. The similarity of the behavior is accounted for by its functionality. But there are other accompaniments of human love that are less functionally relevant: characteristic bodily feelings, modes of expression peculiar to humans and closely related species, genetically induced emotional associations, common learning patterns. Some of these may be caused by the hormone secreted when the mother sees its child, or by some other functionally irrelevant characteristic: for example, the location of the neurological substrate of maternal feelings within the brain. So these accompaniments might be there, not because they are functionally relevant, but because they are caused by the hormone in question. Now, homologues are more likely, as a result of their common origins, to share functionally irrelevant characteristics. So, where functionally irrelevant characteristics are relevant to demarcating kinds, the kinds are more likely to be based on homologies.

Suppose that the spider's behavior, though just as advanced in evolutionary terms as human mother love, is caused by a different hormone and neurophysiologically realized in a different way. This might have the consequence that the spider has feelings that are qualitatively quite different from the human, even though the feelings are functionally the same. Suppose that maternality is differently wired up with respect to the spider's brain. This might result in differences of expression. Another point: suppose that human mother love is shaped in a way relevant to human evolutionary history and spider maternality correspondingly so. This might mean that while the functional characteristics of the emotions were comparable, since the salient ecological challenges are comparable, many side manifestations would be different. For example, it might be that spiders cease to feel for their young when they reach a certain age, and care for all little spiders without differentiation until then. Humans, on the other hand, might continue to care for their offspring forever but not for all little humans undifferentiatedly. Humans might care for their grandchildren, spiders might not. Humans might get sentimental when they see pictures of small children, spiders might not. Would these differences not count as evidence that these are different mental states, even though comparable with regard to a central function? It is reasonable to think so. But this means that, to some extent, functionally irrelevant or functionally peripheral characteristics are important to the identification of emotions or feelings as the same or different. This provides evidence that the emotions are indeed homologies.

Opponents of functionalism have long claimed, largely on grounds of intuition and ordinary language, that irrelevant accompaniments of various mental states, including the emotions, are important to their identification. "Neurophilosophers," for quite different sorts of reasons, make the same sort of claim: for instance, they

COMMENTS AND CRITICISM

TAKING THE MEASURE OF DOOM

Kolmogorov's 0-1 Law states that, if $X_1, X_2, \ldots$ is an infinite sequence of independent random variables, and A is a tail event[1] (that is, an event that is independent of any finite subset of the X_i), then the probability of A is either 0 or 1. In "Could the Probability of Doom Be Zero or One?" Martin H. Krieger[2] argues that, given a certain "conservative premise" about human decisions, this theorem may be applied to actual sequences of choices and events, and, consequently, that our own doom has a probability of 0 or 1.[3] We shall argue that Krieger's paper exhibits serious misapprehensions about both the meaning and the degree of rigor of the conditions that must be satisfied for the mathematical theorem that he invokes to obtain.

Three conditions are stipulated in the statement of Kolmogorov's 0-1 Law: (1) the random variables (that is, our choices) must be *independent*; (2) the event A (that is, doom) must be a *tail event*; and (3) there must be an *infinite* sequence of random variables.[4] We shall consider each of these conditions in turn and examine Krieger's reasons for thinking they might apply to sequences of human action and human doom.

I. INDEPENDENT RANDOM VARIABLES

Krieger introduces the notion of a sequence of independent random variables with the "archetypal example" of the sequence of heads and tails obtained by tossing a fair coin.[5] He then asserts that:

---

[1] We enlarge upon the definition of tail events in section II below. For a formal definition, and technical details, see Patrick Billingsley, *Probability and Measure* (New York: Wiley, 1986, 2d ed.), pp. 57, 295.

[2] This JOURNAL, XCII, 7 (July 1995): 382-87.

[3] Although Krieger withdraws from asserting that the theorem is true for any particular human actions, he does maintain the possibility that it could obtain for an actual sequence of actions.

[4] There is also a technical condition that the doom under consideration be a measurable event, that is, an event for which a probability is defined (cf. Billingsley).

[5] Curiously, Krieger gives the example of a *finite* sequence (of length N). Subsequently, he defines the probability of doom as "the suitably weighted fraction of all life's paths, the ones that lead to doom," where "the probability of any single path...is just the product of the probability of each of the choices" (384–85). A more relevant example would be an *infinite* sequence of coin tosses, where any given sequence has probability 0 and any (measurable) subset of the (uncountably many) possible outcomes has a corresponding probability. Krieger's informal definition in terms of fractions and finite sequences obscures the fact that the 0-1 law is not a theorem of elementary or informal probability theory.

0022-362X/98/9503/133–41

> If we have a set of possible strategies and at each choice point we pick out one by means of a die, then each chosen strategy is independent of the others (382).

Notably, it is not a sufficient condition for strategies to be independent that they be random, an important distinction which Krieger's discussion does not make fully clear. For instance, suppose we make a sequence of bets, where each bet is a dollar more or a dollar less than the preceding bet, depending on the toss of a coin. Each bet is random, but the bets are not independent, since the possible values for each bet depend upon the value of the previous bet. What is required for a sequence of random choices to be independent is that the distribution of each random choice not be affected by the values of earlier or later choices.[6]

Krieger comes unstuck when he moves from independent random events to human actions. Prima facie, it seems patently false that our actions are either random or independent of each other, because we make decisions for reasons, formulate plans, and act upon them, and we are influenced both by what we observe and by memories of our previous actions. Krieger here introduces his "conservative premise." He proposes that the information upon which we act could consist in "random processes in which there is no correlation between events" (382), or *noise*.

> A traditional conservative might well argue that while most of [the] time we think we know what we are doing, we are actually acting on noise, especially in the synoptic or comprehensive realm (and hence the conservative's doubts about rationality or planning). The point here is that, although 'no coherent planning' may seem like a very strong and even outrageous assumption, perhaps it is true (at least in effect) (387).

Krieger does not tell us whether he is a traditional conservative, but he is at least engaged in the task of showing that such a view is defensible—that is, that despite appearances, the "information" that informs our choices is mere random data. Krieger recognizes that, if the 0-1 law is to apply, he must show not only that the information upon which we act could be bogus, but also that our apparently rational and thought-out activities could, in fact, be random behavior, or that we could be "noise deciders." He believes that this follows staightforwardly from the traditional conservative's assumption, such that to be noise deciders it is sufficient that "our choosing nontrivially depended on what we took to be information" (383), that is, that every choice be influenced by noise. But this condition is too weak to guarantee that our choices are entirely random. If I am shopping for food in a supermarket, my choice is nontrivially dependent on what is on the shelves, in the freezer, and so on, but the available stock does not determine which particular item I pick. Even if the information that I have about what food is available is noise—if all the labels on the cans and packages giving details about contents, weights, sizes, and so on are random—there is still room for me to make choices based on my independent ideas about how to dine in style. I could, for example, buy only a few items and rely on my culinary expertise to make the best of the ingredients I purchase; alternatively, I could adopt the strategy of buying very large quantities of food to be fairly sure of avoiding a gustatory disaster. My choices are not entirely random: although what I buy has a random element, I have an influence over the *distribution* of the corresponding random variable. If we are to be noise deciders, then more is required than that our choices be nontrivially determined by noise.

---

[6] See Billingsley for a formal statement.

Perhaps Krieger's traditional conservative is alive to this type of difficulty, and would argue that noisy information pervades every stage of the decision process. But even in this case, there is a distinction to be made in the degree to which our decisions may be determined by noise, and this effects the scope of Krieger's argument. Consider the following three (nonexhaustive) possibilities:

(1) Our decisions are entirely random and are unaffected by our natures as human beings or our individual characters.

(2) Our choices are influenced by those characteristics which distinguish us as human beings, but are not contingent upon individual characters.

(3) The information upon which we make decisions is random, but our choices may be individual and idiosyncratic according to our personal preferences.

To conclude from Krieger's argument that two agents X and Y have the same probability of doom, it is necessary not only that X and Y make random choices but also that the random variables X_1, X_2,... (corresponding to successive choices) have the same distribution for both X and Y. So for each degree of freedom we have distinguished, (1)-(3), there corresponds a possible interpretation of Krieger's argument:

(4) The probability that X faces a given doom is either 0 or 1, independently of what X does and who or what X is.

(5) The probability that X faces a given doom is either 0 or 1, depending upon whether X is a human being, though not upon what sort of human being X is.

(6) The probability that X faces a given doom is either 0 or 1, depending upon who X is.

Krieger does not indicate which of these interpretations he intends. It seems that the second or third arguments are preferable, however, because in these cases the decisions that we make and the probabilities of our dooms are either up to us as individuals or are determined by those features which make us human. So even if we accept that certain sequences of our actions are doomed, we can reject the "apocalyptic" conclusion—for which the traditional conservative, and perhaps also Krieger, seem to be vying—that doomed events are entirely beyond our control. Suppose I am the captain of an oil tanker. Even if my navigation of the ship across the Atlantic is doomed with probability 1 to end in my running the ship onto the rocks, it may nevertheless be the case that had someone else been in charge, or had my employers trained me better, or had I been less reckless, and so on, the disaster would not have happened.

Of the three interpretations of Krieger's argument, the final one is the weakest. But even in this case, the conditions are extraordinarily difficult to meet. For example, an agent's choices need not be independent if the agent were to undergo any changes of character or develop any preferences for a given type of action. The agent's prior experience could not also enter into the decision making. Krieger suggests that much trading in security markets may be based upon noise. But in this context, traders would call upon their memories of both the past behavior of particular securities and of securities in general in making their decisions (which is seen most obviously when the market crashes). Consequently, their decisions are not in practice independent, even if the incoming information regarding the state of the market is noise. Indeed, it seems difficult to imagine any naturally occurring sequence of human decisions which satisfies the independence condition of Kolmogorov's 0-1 Law.

II. THE INDEPENDENCE OF DOOM

For Kolmogorov's Law to apply to our choices, not only must our choices be independent of each other but also the doom under consideration must be a tail event, that is, an event that is *logically* independent of *any* finite set of our choices. In other words, the occurrence of doom cannot be changed by altering finitely many of our decisions. For example, in an infinite sequence of coin tosses, the event that infinitely many tosses come up heads is a tail event.

In order to defend the assertion that doom is a tail event, Krieger distinguishes between three positions, which he labels as *apocalypti-*

cist, *marginalist*, and *modest skeptic*. Apocalypticists believe that "the cumulative possibility of doom in time is one" (385); they are, in other words, sure that doom will happen. Marginalists are more hopeful and believe "the probability of doom is zero": they are confident that we shall be able "to wiggle our way out somehow" (385). These views are contrasted with modest skeptics, who believe that the future is never determined: "it is never surely too late to save ourselves, and the straight and narrow is not that restrictive. Also, it is never surely safe; we might slide down a dooming path we could not avoid, and there may well be more than a few of these paths and their measure need not be zero" (385). (One might alternatively take the view that our actions are not doomed at present, but may become so at some future time.)

Since modest skepticism is a position to which many of us might well be inclined, it is important that Krieger be able to give an account of it, particularly since the other options he gives, apocalypticism and marginalism, are taken by people who already accept the conclusion of his argument that the probability of doom is either 0 or 1.

Krieger argues:

> For a modest skeptic, it is never surely too late, never surely safe. If we conceive of a path as having an unending number of steps or choices (whatever their size), then at step *n* we could not know whether we are surely too late or we are surely safe; for we are not at all sure. Further steps can alter our path's destination. The final state is, in effect, independent (and I shall model this by statistical independence) of any given finite set of steps (385–86).

The modest skeptic believes that, after a finite number of choices have been made, neither doom nor the absence of doom is certain. Krieger's model of this theory contains the premise that the occurrence of doom is independent of any finite set of choices.[7] But this is a huge additional assumption not made by the modest skeptic. If an event, A, is independent of any finite subset of the random variables, X_1, X_2,..., then the probability of A must remain the same if we condition on any finite set of values:

---

[7] Krieger is wrong when he defines a tail event as "statistically independent of any finite number of the steps we make on the way to it" (384). The independence must be logical: there is no mention of probability in the definition of a tail event (see Billingsley, example 22.2). Statistical independence is a consequence of logical independence and the other conditions of the 0-1 law. As we argue in this section, however, Krieger does not even address the requirement of statistical independence.

$$P(A \mid X_1 = x_1, \ldots, X_n = x_n) = P(A)$$

In other words, the first n choices do not affect the probability of doom. So, whereas Krieger's model requires that the probability of doom after the first n choices is the same as it was initially, whatever the choices we have made, the modest skeptic claims merely that making such a sequence of choices cannot seal our fate either way. Indeed, the modest skeptic is entitled to believe that the probability of doom could be altered by individual choices, whereas Krieger's model cannot allow this possibility. Krieger therefore fails to give a satisfactory account of modest skepticism.

III. INFINITE SEQUENCES OF CHOICES

A particular doubt about the applicability of Kolmogorov's Law to human choices arises from the third condition, which asserts that the sequence of random choices must be infinite. It seems that an infinite number of choices should take an infinite period of time to complete. There are two strategies that we might use to avoid this problem: we could analyze our behavior as a Zeno-like supertask consisting in an infinite sequence of actions with ever shorter duration, or we could consider situations involving an infinite number of agents.

The difficulty in analyzing our behavior as a supertask is that none of the Zeno paradoxes and related supertasks consists in random choices. Achilles, in his race against the tortoise, acts determinately in successively covering the remaining portions of the race course. Consequently, the infinite sequence of actions Achilles performs before crossing the finishing line (the "doom" event) does not satisfy the conditions of Kolmogorov's 0-1 Law. But we could consider the following example. Suppose that Achilles, thoroughly inebriated after drowning his sorrows because of his failure to win his race with the tortoise, decides to run over the course once more. He must first cover the initial half of the course, then the next quarter, then the next eighth, and so on. He can cover each of the infinite sequence of lengths in an ever shorter time, namely, half a minute, then a quarter of a minute, then an eighth of a minute, and so on. At the beginning of each time period, however, he may find himself too intoxicated to move, so that whether he runs the next segment of the course or stays where he is becomes a random event.[8] The "doom" of this supertask is constituted by Achilles succeeding in staggering to the finishing line, and can be shown to be independent of any finite sequence of steps he may take to get there. We can therefore apply Kolmogorov's 0-1 Law, and conclude that the probability that the drunken Achilles completes the course is either 0 or 1.

---

[8] This is like Paul Benacerraf's staccato runner. By appropriate modification, we can avoid problems of unbounded speed.

A second way of avoiding infinite time is to have an infinite number of agents making choices (it is not necessary that the infinite sequence consist of *sequential* choices: although they must be ordered as X_1, X_2,..., they need not occur in any particular temporal order). Suppose, for example, that a general election is held with an infinitely large electorate. The political information given to the electors to make their decisions is noise, as is the information provided by the polls, the media, and so on. Furthermore, voters make their choices independently of one another. Such an election can be modeled by an infinite sequence of random variables: the "doom" event of a particular candidate's receiving an infinite number of votes then has probability 0 or 1 (this event is clearly independent of the votes cast by any finite set of electors).

Whereas the drunken Achilles and the infinite election are examples for which Krieger's thoughts are pertinent, it is doubtful whether either example could conceivably occur in the real world. Thus, in addition to an implausible hypothesis about the nature of information and the nature of doom (that is, that it is independent of any finite set of decisions preceding it), Krieger is encumbered with the defense of an additional hypothesis: that there are either an infinite number of agents involved in the decision process, or that a single agent might make an infinite number of independent, distinguishable decisions in a finite period of time.[9]

A possibly more attractive line of argument open to Krieger is that there could be an infinite sequence of random variables, where only a finite number correspond to choices and the rest correspond to random physical events. Although Krieger assumes throughout that the infinite sequence is constituted by human choices alone, the mathematical argument does not require that the random events be the result of human choice: random physical events will do just as well. So, without involving himself in the difficulties that arise from there being an infinite number of human choices, Krieger could instead defend the possibility of an infinite number of random events.

---

[9] Krieger implicitly suggests the latter of these two options as a possibility when he proposes that a modest skeptic might "conceive of a path as having an unending number of steps or choices (whatever their size)" (385). Practical difficulties aside, a subdivision of a sequence of independent random choices into a sequence of very small random steps does not generally produce a sequence of independent choices.

Unfortunately, such a sequence would not support his larger argument. If only finitely many of the random events are human choices and doom is independent of any finite set of events, then doom is independent of the set of human choices: by assumption, and without any appeal to Kolmogorov's Law, doom is not contingent upon our actions, which are thereby irrelevant. Doom is a natural event, and it is impossible for us to avert or cause it by our actions.

IV. CONCLUSION

The insurmountable difficulty for any application of Kolmogorov's Law is the stringent character of the conditions. Many mathematical models are useful even when their assumptions are not quite satisfied; for instance, Newtonian mechanics offers an approximate but useful model for the physical world. But the same cannot be said for Kolmogorov's Law. Krieger indicates that the conditions of the 0-1 Law are only approximately satisfied: the final state is "in effect independent (and I shall model this by statistical independence) of any given finite set of steps," we "stipulate" that the choices are represented by independent variables as "is often done in fault-tree and decision-tree analysis" (that is, the *model* of the choices satisfies the 0-1 law, whatever the nature of the choices themselves), the assumption of no coherent planning may be "true (at least in effect)" or at least "most of the time." Indeed, in the conclusion to the main part of his argument, Krieger states

> Now, it may well be the case that this conclusion is actually quite plausible. Doom is a *rare* event and *may well be outweighed* by survival. Or, doom is surely to happen somehow, and we live on borrowed time whose payback period is *long compared to our historical notions* [emphasis added] (386).

But it is not enough for the conditions of the theorem to be approximately satisfied: the sequence of choices must actually be infinite, the random variables must actually be independent, and so on; furthermore, if the conditions are only satisfied approximately, then the conclusion need not be satisfied approximately.[10] In this respect, Krieger's discussion of "error estimates"(385) is extremely misleading, as is his suggestion that "[i]n so far as we are acting on noise" we are in "a very different realm" from traditional analysis (387), since he does not provide an approximate version of the 0-1 law or a theoretical error analysis.

---

[10] For instance, suppose we toss a fair coin an extremely large number N times, and consider the "doom" event A that there is an even number of heads. The tosses form an independent sequence of random variables, and A is statistically independent of any N-1 tosses. But A has probability 1/2. It would be easy to give other examples demonstrating that the conditions of the 0-1 law are not negotiable.

Krieger is no less evasive when he comes to give an account of how all three conditions that we have outlined could obtain. He takes the issue to be one of whether we should "abandon the Kolmogorov context" for some other model of doom, and he suggests, first, that probability estimates of doom given by other models may not be robust if subject to systematic errors, and, second, that we have a general "commitment...to the existence of information" that is "even more curious" than the possibility that Kolmogorov's Law applies (387). But neither of these suggestions is much to the point, because, as we have argued, the issue is quite the reverse: our being noise deciders is only one of an extraordinary collection of circumstances that must hold for Kolmogorov's 0-1 law to have the slightest bearing on the probability of doom. It is possible, of course, that a stronger version of the 0-1 law or an entirely new result could be proved, and this might well turn out to have not only mathematical but philosophical interest. Alternatively, Krieger might describe some human doom for which the conditions of the 0-1 law actually hold. In the absence of either event, however, we must conclude that Krieger's position is untenable.

ALEXANDER D. SCOTT

University College/London

MICHAEL SCOTT

University of Utrecht

BOOK REVIEWS

Kant after Duchamp. THIERRY DE DUVE. Cambridge: Massachusetts Institute of Technology Press, 1996. xv, 484 p. Cloth $39.95.

"Everything [in my art] was becoming conceptual, that is, it depended on things other than the retina," Marcel Duchamp told Pierre Cabanne in a well-known interview of 1987. Insofar as the twentieth century has given art over to an increasingly theoretical, philosophical, or, more broadly speaking, conceptual activity, with what degree of ease can an essentially—dare one say it—visual medium turn itself into what the conceptual artist Joseph Kosuth has called a "purely propositional" medium? Kosuth's demand that art become purely propositional is a claim of art theory voiced as a normative requirement. Yet it is worth noting that even Kosuth's actual conceptual installations are partially at odds with his theoretical claims about their purely conceptual identity, because the very attempt to curtail visuality and "discipline" it into the form of the purely propositional generates its own strange visual insistence. What Kosuth does is to "create space" which exists as an empty vessel composed of objects drained of all ordinary meaning and use. This dematerialization of space/object becomes, like the jar in Tennessee of which the poet Wallace Stevens speaks, strange, uncanny, fascinating, even sublime, and the work remains firmly planted in the realm of Immanuel Kant's aesthetics (if only through its attempt to break out of that realm). As for Duchamp, he remains visually exploratory, ironic, fascinated, and playful, if also suspicious and interrogative.

These remarks about the partial disjunctions between art theory and the meaning or force of the artwork are not incompatible with the highly imaginative, wide-ranging, and ambitious reading of the "archeological shape of modernism" which De Duve offers. Luxuriating in the bric-a-brac of twentieth-century art and culture, his book competes with Charles Baudelaire's famous painter of modern life in its fascination with the "here and there" of modernism. Although De Duve's readings are often excellent, when he finally settles down to the task of formulating and presenting a thesis, the result is as problematical as it is inviting.

The thesis is this. Modern art has been shaped by an antinomy of taste, which only a return to the pages of Kant's *Critique of Judgment* will "solve." What is the antinomy? On the one hand, those like Jackson Pollock working innovatively in traditional mediums of art

0022-362X/98/9503/142–48

(painting, sculpture, and so on) have produced "works" for which the key question surrounding their reception has been: 'Is this a painting (sculpture, and so on)?' For the artworld to have come to understand that Pollock's productions were paintings was for it to have found a way to appreciate the aesthetic values in them. Thus the question 'Is this a work of art?' is answered affirmatively in Pollock's case only when two prior questions—'Is this a painting?' and 'Does it have aesthetic value?'—are jointly addressed. On the other side of the "archeological divide," conceptual art and theory have engendered objects for which a wholly different set of questions has been in play. The basic question raised by a Duchampian *Readymade*, De Duve thinks, has neither been that of the kind of artwork it is, nor the related question of the kind of aesthetic value it has. Instead, these unruly works have occasioned the basic question 'Is it art?', which has been, in their cases, answered independently of recourse to questions either of kind or of aesthetic value. Putting the two sides together, '*X* is art' both is and is not tied to '*X* is of kind *p*' and '*X* has aesthetic value *v*'.

There is no doubt that modernism has exhibited profound disagreements over the roles of the aesthetic and of traditional mediums in art. Nor is there any doubt that the propositions which artists have pronounced or inscribed in their manifestoes, and which critics have written in support or rejection of their work, generate a range of inconsistencies. But my reason for opening with disjunctions between theory and visuality in the work of conceptual artists like Kosuth is that such rifts point to an inability to draw inferences from theoretical pronouncement to the meaning or force of the artwork itself. De Duve seems to rest easy, however, in the assurance that such inferences can be drawn. Because of this assurance that what is said and what an object or installation "means" are of a piece, he never properly engages the issue of how theory in modern art plays a role in the formation of larger art practices, and how these art practices, in turn, inform or fail to inform the meaning or force of the "work of art." De Duve's book cannot then be said to provide an "archeology of modernism." For the project of an archeology, as Michel Foucault defines it, is that of analytically unfolding a "discursive formation" according to which rules for the production of new truths and new forms of visuality are unearthed in the "archive" (here the modernist archive)—and also of showing how these rules serve to articulate a system of modern power. Neither rules nor power are unearthed, in spite of De Duve's Foucauldian presumption.

De Duve's philosophical claim is, we have seen, to have discovered in the archeological space of modernism a specifically Kantian antinomy. Yet consider what would be at stake in making good on this claim. A Kantian antinomy is an artifact of reason's passing beyond its limits into the making of claims whose defects consist in the fact that they or their contraries can be argued with equal "conviction." Moreover, the very nature of an antinomy is that it admits of a Kantian solution, which, having modestly tailored reason back to its limits, can allow the conduct of reason (that is, of cognitive judgments) to go on in a legitimate way. Indeed, the very role of antinomies of taste as they inform Kant's *Critique of Judgment* (where taste admits of no cognitive judgments) is, for some commentators, very obscure and, for others, highly suspect. Thus, for De Duve to argue his case for an antinomy of taste in the field of modernism is for him to show, first, that antinomies of taste are not suspect and, second, that the inconsistency he discovers in taste can, in the manner of any antinomy, admit of a "Kantian solution" which will allow the conduct of judgment to proceed legitimately once limits are set on that form of judgment.

Let us pursue how De Duve goes about solving the "antinomy." The "resolution" is nowhere very clear, but it seems to have two stages. First, De Duve claims that the judgment '*X* is art' is already a Kantian judgment of taste, even when Duchamp or Kosuth claims to fracture it from any judgment about aesthetics. What this amounts to is that the judgment '*X* is art', even when used by a Kosuth, has two indelibly Kantian features. First, in accord with Kant's first moment of the beautiful, it is a judgment rooted in the experience of delight, or at any rate in the experience of "feeling." Second, in accord with Kant's second moment of the beautiful, the judgment is articulated in the "universal voice," with the subject's placing herself in the position of an exemplar for all humanity (or at any rate for all players in the artworld of the West) when she judges.

By assimilating '*X* is art' already of its own accord to the realm of judgments of taste, the antinomy between art and aesthetics is putatively removed. But what is the use of this, given that De Duve hardly tells us very much about what taste is? While Kant's first moment of the beautiful specified that the judgment of taste is a very particular (if highly obscure) kind of feeling, namely, a delight taken in the object apart from all interests, De Duve merely tells us that some (any?) kind of feeling is involved. There is not much aesthetics in this claim, since anyone passionate about anything will, by definition, have feelings about the thing in question. As for the second moment, no doubt each player will presume to speak in the universal voice; but then, does not every child of the Enlightenment? Moreover, whether or not '*X* is art' is reclassified as a judgment of taste, we are still left with the conflict between the two sides of the "antinomy" intact. Kosuth will still claim that his art should be (is) free of the aesthetic, while Clement Greenberg (in one mood), Michael Fried (in every mood), and Pollock in his painting will claim, with equal feeling, that Kosuth's position unhinges and degrades the very concept of art.

De Duve's second strategy involves the unfolding of three claims that are meant to "give the semantics of" the judgment (sentence) '*X* is art' in a way that, again, resolves the antinomy between it and judgments about aesthetic value. First, it is stated that 'art' is a proper name. It is possible to dispense with this claim quickly, because it is not—any more than 'computer', 'tool', or 'video player' is. While such words can become names under special circumstances (as when someone baptizes their computer 'Computer' or their dog 'Dog'), such a joke would happen in the field of art only occasionally. That it did happen with Duchamp, who called his limp sculptural "phallus" *Objet-dard*, is a sign of his own particular aesthetic style, not of anything generalizable into a semantic theory of art. Second, De Duve suggests in a long discussion of Saul Kripke, which constitutes the least successful part of the book, that the term 'art' is a rigid designator. Again, it is not, any more than 'video player' is. Rigid designation implies that the reference of a term is set for all possible worlds. To use Kripke's famous example, the term 'meter' is defined by that peculiar readymade, the standard meter bar in Paris. While our beliefs about metrical measurement might change, a wedge is drawn between meaning and belief with respect to the term 'meter', for the "meaning" of the term is set by its reference to the thing in Paris. This picture of belief, meaning, and reference is highly inadequate for a term like 'art'; for as often as not, modernist debates center around the *references* of the term (with castigations of the 'It is not art' variety flying about the museum like so many epithets).

Third, De Duve culls from Duchamp the thought that '*X* is art' has a nominalist semantics. Rather than engage in a lengthy discussion of what Duchamp—or De Duve—might mean by this, I prefer to suggest that he is groping for the following thought, again one associated with Kant: our position of judging some object to be art is a radically particularist position. We may indeed judge it comparatively, in terms of its complex connections (aesthetic, thematic, institutional, historical, and so on) to other objects that fall under the concept of art. But each object defines a distinct field of such comparisons on its own terms. Now, until the particularist position can be adequately squared with ordinary art practices, in which comparisons are imposed on objects as well as occasioned by them, this position exists at the level of a mere proposal. So we are back again to the conflict between visions of art with which we started, for there has been no Kantian solution.

Let me therefore pursue the question of a possible antinomy in the field of art practices from a different perspective, a more archeological one. I have suggested that in spite of what Kosuth says, his installations never exit the field of the aesthetic. I would add that it is in the space of misalignment between word and object that his modernist practice lives. Pursuing the importance of this space for modernist practices generally, perhaps we might find that "antinomial" claims when uttered by the likes of Greenberg, and misalignments between word and object when engendered by the likes of Duchamp, are precisely what lend this region of modernism its pistons of energy. Consider, that is, the thought that the role of theory in modernist art practices is not—or not reducible to—that of the production of interpretive truths about the modernist art object. As I have argued in my own previous work on modernism,[1] I think this is the case.

Since De Duve provides an excellent description of Josef Beuys, let us pursue this issue of force and energy vis-à-vis that artist's writings. Beuys claimed that everyone could become an artist. Now, when he claimed this he knew very well that it was a utopian claim, but what he failed to grasp was its ultimate incoherence. His is certainly not the claim that everyone can become capable of making aesthetic objects in plastic mediums—which is plainly false (I cannot draw to save my life, nor sculpt either). His is a claim expressing a vision of the total transformation of art and artists. Whatever an artist does is now art, and there are no constraints on what an artist is. We exist in a phase of complete liberation. With such a trumpet call in the direction of transcendence, 'artist' (and by extension 'art') becomes a concept without criteria, and such a concept is not in the end a concept at all but rather the utopian image of one. It is central to Beuys's rhetoric that the concept of art is therefore rendered *sublime*, as sublime as Kosuth's object/space is.

Yet—and hence comes the apparent antinomy—in another voice Beuys does speak to what this new vision of art will be. For "everybody to become an artist" will be for each and every person to learn the work of reenactment of the past and its purgation through the imagination—a cultural, specifically German, sort of work, the work of the phoenix returning to its mythological origins. R. G. Collingwood had prefigured Beuys without realizing it when he argued in *The Principles of Art*[1] that the real work of art is what takes place in the human imagination when an emotion is worked through or "expressed" in it; and Collingwood was happy to accept that this relegated the art object to a merely supplementary status. Beuys may also be said to provide an aesthetic of expression which favors process over product. Finally, we may find in Beuys's writings and visual objects still another aesthetic: that of a member of the avant-garde movement Fluxus and originator of the "Happening" who wishes to turn humanity into an active participant in spectacle—the inventor of a kind of counterspectacle to those of his youth in the *Hitler Jugend*.

---

[1] *Making Theory/Constructing Art: On the Authority of the Avant-Garde* (Chicago: University Press, 1993).

These voices are hardly consistent. But do they compute to an antinomy which functions in "the usual Kantian way"? Is the point to "solve" it, thus placing judgment on a legitimate footing? In some sense, "yes": the work of critique would do that. But more importantly, Beuys's "antinomial" use of language must be understood *to have an essentially aesthetic function*; it is part of what defines Beuys's style—an avant-garde style oriented around the aesthetics of the sublime, also around the aesthetics of reenactment and the spectacle. As such, there is no resolution to it that is desired, merely an understanding of the role of Beuys's concepts in a utopian art process. The "theory" is itself part of what drives the avant-garde practice, for it turns sentences about art into *art* by making them sublime, a sublimity requiring incoherence, therefore unconceptualizability. We resolve this antinomy only insofar as we must point to its incoherence at the level of theoretical judgment, a form of incoherence in fact attached to all utopian thinking per se; otherwise, we let it be. If Beuys is to be criticized, it is because his use of concepts and organization of spectacles aestheticizes politics in "the usual German Romantic way."

My point is that when antinomies do arise in the course of radical art theory, we must understand the role of that theory in the texture of the art practice to be that of a "tool in the language game" (Ludwig Wittgenstein) with contextually defined purposes. The field of what looks like antinomial positions between artists and critics is, in the modern period, both an aesthetic occasion and a public sphere in which debates between such positions—positions articulated in the light of art-historical and institutional training, but also in the heat of engagement—may take place. This field, through which a number of different subjects (artists, critics, viewers, and others) may exercise their autonomy, may or may not be "resolvable," depending on the positions taken. It is the *force* of positions as well as their interpretive meaning which is central to this sphere. May there long be art of sufficient quality, originality, and character—art free from boredom, complacency, competition with television, or political correctness—to allow us to rise to the occasion of public debate and the drama of envisionment, as Greenberg, Duchamp, Kosuth, and now De Duve have done.

---

[1] New York: Oxford, 1938.

DANIEL HERWITZ

University of Natal/South Africa

The Unity of Reason: Essays in Kant's Philosophy. DIETER HENRICH. Edited by Richard Velkley. Translated by Guenter Zoeller, Manfred Kuehn, Louis Hunt, and Jeffrey Edwards. Cambridge: Harvard University Press, 1994. 249 p. Cloth $45.00.

The Course of Remembrance and Other Essays on Hölderlin. DIETER HENRICH. Edited by Eckart Förster. Translated by Abraham Anderson and Taylor Carman. Stanford: Stanford University Press, 1996. xi + 302 p. Cloth $45.00.

The Unity of Reason (UR) contains translations of four of Dieter Henrich's most important essays on Immanuel Kant, including "Identity and Objectivity." *The Course of Remembrance* (CR) comprises translations of Henrich's work on the philosophical importance of the poet Friedrich Hölderlin. Together they are a sustained meditation on the central issues and themes of classical German philosophy which show to great effect Henrich's magisterial historical grasp and analytical skill. His treatment in these books is rooted both in (1) a particular methodological approach to the question of the relation of philosophical historiography to contemporary philosophical interests and (2) a conviction that, appropriately excavated from their local historical context, some of the substantive philosophical positions developed in German idealism remain relevant. This methodological approach, which Henrich calls *argumentative reconstruction,* first identifies concepts central to the philosophical theory in question which admit of various constructions consistent with the general commitments of the theory, then assigns these concepts fixed values allowable under the theory, thereby generating different reconstructed models of the theory (UR 124). Questions of which model a thinker in fact favored and which she should have favored given her overall views are, in principle, separate.

I

Henrich's monograph-length study of the transcendental deduction in the *Critique of Pure Reason,* "Identity and Objectivity," is the heart of *The Unity of Reason.* It would be difficult to overestimate the importance of this complex and provocative essay. I give here only the most schematic account of its arguments. Henrich famously contends that there are three main lines of possible argument which can be extracted from the deduction which attempt to make good on Kant's claim that experience requires synthesis according to the a

0022-362X/98/9503/149–59

priori concepts of the understanding. Arguments from the structure of judgment are unsuccessful: judgments of the categorical subject-predicate type with which Kant is concerned do require internally complex objects, but leave inter-object experience ungrounded. Arguments from the unity of the subject fare a bit better, but they, too, come up short. Henrich asserts that only the argument that takes as its premise the subject's a priori knowledge of its "moderate," as opposed to "strict," numerical identity through temporally successive representational states is ultimately successful (UR 179).[1] Moderate numerical identity allows that a thing is the same thing so long as it exhibits continuity through some, "but never total, change[s] of its states" (UR 180). Once we make this qualification of the sense of identity involved in the premise of the deduction, the Kantian argument for the necessity of the categories from self-consciousness falls readily into place. Kant holds that we have a priori and certain knowledge of ourselves. Part of such knowledge would have to be knowledge of the identity of self. But that can be thought only in terms of identity of self remaining constant over alteration in representational states. So, the subject must "have in advance of all experience a knowledge of what it means for it to pass from one representational state to another" (UR 187). Knowing this requires knowing the *ways* such transitions must take place. This means, thinks Kant, that transitions between representational states are due to rule-governed synthesis. These rules are the categories.

Henrich's analysis of the deduction is extraordinarily rich and suggestive and, as such, has been a spur to both admiration and criticism. The clearest case of the former is Paul Guyer's approach to the deduction which adopts Henrich's insights as a general scheme within which further possibilities are identified. Yet, as powerful as Henrich's analysis is in the details, it raises questions and reveals tensions in his approach to the deduction. This tension can be viewed as one which results from conflicting allegiances Henrich seems to have to elements of the first (A) and second (B) edition deductions. Elsewhere, Henrich[2] has taken pains to emphasize the B deduction's preeminence by stressing that to show that *all* (and not just *any*) representation must be subject to the categories, Kant must demonstrate that intuitions themselves are products of categorial synthesis—namely, that even apprehension is conceptually guided—a position taken expressly only in the second edition (B159).[3] Moreover, the reconstruction that Henrich offers here crucially appeals to the *temporal* succession of representational states over which the identity of the subject ranges, and thus naturally recalls the second edition's integration of the constraints levied upon experience by the forms of intuition into an account of conceptual synthesis. But Henrich's analysis bypasses the B deduction almost entirely in favor of the A deduction, where Henrich locates the primary textual support for his reconstruction of the argument from moderate identity. Kant offers as the premise of the deduction: "we are conscious *a priori* of the complete identity of the self in respect of all representations which can ever belong to our knowledge, as being a necessary condition of the possibility of all presentations" (A116). Henrich reads the object of which we are conscious a priori to be specified by the clause 'the complete identity of the self', and the final clause 'being a necessary condition...' as a further description of the identity. This statement of the apperception principle is the synthetic proposition: whenever I am conscious of my identity as a mind containing various representations, it must be the case that I am also conscious that the representations are subject to the condition of my identity—their synthesis according to rules (the categories). This stands in clear opposition to the analytic statement of the principle in the B deduction, favored by commentators like Henry Allison; that is, the hypothetical claim that, if I have representations in a unified intuition (ones that have been synthesized), then these representations are subject to the general conditions placed upon representations for belonging to one consciousness (B131-32, 134, 138). Unless one is willing to argue that the synthetic statement of the principle is also implicit in the second edition (and I think this implausible),[4] Henrich's exclusive attention to the A deduction is given at the expense of what Kant says in the second edition. Some may also balk at Henrich's ascription to Kant of the thesis that the consciousness of identity must be one of "Cartesian certainty" (UR 186-87). The idea that the subject must be certain that all that is given in intuition must conform to the subject's capacity to know its identity may seem an unwelcome and uncharacteristic nod to René Descartes. Finally, as Guyer[5] has pointed out, Henrich has voiced his strong opposition, both here and elsewhere, to Kant's "constitution talk." But the very heart of the A deduction is the psychologistic analysis of synthesis that commits Kant most expressly to the idea that the world of objects is, in some sense, "constituted" by judgment.

---

[1] The moderate conception of the identity that Henrich adopts, first urged by C. A. Crusius against G. W. Leibniz, argues that, since the concept of identity is concerned to explicate what it is to be the same thing given changes in the state of that thing, definition of identity in terms that excludes the possibility of alteration of a self-same thing (as the doctrine of the identity of indiscernables does) is inappropriate.

[2] "The Proof Structure of the Transcendental Deduction," *Review of Metaphysics*, XXII (1969): 640-59.

---

[3] Citation to Kant's *Critique of Pure Reason* uses the customary A/B format to indicate its first and second editions. All other citation of Kant's work is to the standard Akademie edition by volume and page number.

[4] There have been attempts to locate subterranean statements of a synthetic apperception principle in the second edition. See Paul Guyer, "Kant on Apperception and A Priori Synthesis," *American Philosophical Quarterly*, XVII (1980): 205-12.

"The Concept of Moral Insight and Kant's Doctrine of the Fact of Reason" and "The Ethics of Autonomy" are companion pieces, in which Henrich discusses, respectively, Kant's unsuccessful arguments for the validity of the moral law in the so-called "silent decade" and the importance of the mature Kantian account of autonomy for later idealism. Kant's search for a deduction of the validity of the moral law is conditioned by his attempt to explain what Henrich calls *moral insight* (UR 57), a term meant to canvass two properties of moral judgment: its objective validity and its unique intimacy to action. An adequate account of moral insight must include an understanding of what distinguishes moral from scientific knowledge (its intimacy to action), while insisting upon the epistemological character of moral judgment (its objectivity). Now, Kant accepts the *idée reçue* that the intellect alone has no power to motivate action: motivation requires affect. Moral insight, therefore, must have an affective component (UR 73) and Kant's early reflections on the nature of moral judgment are attempts to identify the affect appropriate to moral motivation. Henrich identifies two complementary lines of argument for the validity of the moral law in Kant's precritical writings, both involving proofs of the moral law from the general structure of theoretical reason. Struck by the analogy between (1) theoretical reason's striving toward unity and concomitant vocation of categorizing and systematizing manifolds of intuition or concepts and (2) the moral imperative's function of unifying the will, Kant considers the argument that the subsumptive feature of moral judgment is a special case of a more general theoretical capacity for generating maximal coherence among thoughts (UR 74-75) and that the unity of affect and reason is a similar pleasure in system. A variation on this line is also attempted: freedom without lawfulness is unpleasant because it destroys the order so prized by reason. But lost in both attempts is what makes moral and theoretical judgment essentially different. The demand of morality to do what must be done, come what may (its categorical nature in the face of potentially countervailing inclination), cannot be adequately captured by the formal striving for systematicity in the will. I do what I must because I know that I must, not because I deem the action to promote the maximal systematicity of my thoughts and desires (UR 76).[6] Henrich distinguishes these "direct" attempts at arriving at an adequate account of moral insight from one "indirect" attempt which uses transcendental freedom as a "mediating idea" between the understanding and the good (UR 75). Here, the idea is to deduce moral insight from freedom as a presupposition for the possibility of ethics. This, of course, is the same path which Kant chooses in *The Groundwork for the Metaphysics of Morals*, Section III, and which he later renounces.

---

[5] "Review of Dieter Henrich: *Identität und Objektivität*," this JOURNAL, LXXVI, 3 (March 1979): 151-67.

The problem of accommodating affect in a theory of moral judgment remains central to Kant's mature moral philosophy and informs his account of autonomy. For a will to be autonomous it must both be able to know what it wills, prescinding from any knowledge about inclination (what Henrich calls *autognosy*), and have, native to itself, the "power to effect actions which take place solely because they are rational" (UR 95). This is the fashion in which Kant's mature moral thought attempts to capture the cognitive and conative elements of moral insight. As to the affective element of the moral enterprise, Kant argues for a sui generis nonpathological feeling of respect for the moral law as motivationally efficient. This feeling is itself rational and thus is not causally dependent on the empirical vagaries of sensible experience. The concept of respect for the moral law, therefore, is specifically crafted to yield a correlate to empirical feeling for pure reason. But Kant's solution to the problem of pure rational moral motivation comes at a price. This is evident when we consider the phenomenological character of respect. The felt quality of respect is explicated by Kant in virtue of certain negative relations that accrue between moral rationality and sensible feeling. That is, although the feeling of respect for the moral law is sui generis, its content depends crucially upon conflicts between our rational and sensible natures. In respect, we feel the elevation of our rational natures, but they are elevated only at the expense of striking down our

[6] Kant does not entirely abandon this parallel between theoretical and practical reason's systematicity. He continues to think that practical reason seeks a comprehensive system (1) of its own laws and (2) of a maximally systematic ordering of theoretical and practical laws in one world. This is the import of the twin notions of a "kingdom of ends" (A808/B836; A815/B843; Ak. IV, 433, 436n.) and the *summum bonum*; see also A807-19/B836-47. What Kant abandons is the idea that systematicity can play a constitutive role in moral psychology.

sensibility. This has a compensatory feel to it, reflected in the dual quality of the feeling of respect, as at once pain and pleasure. As Henrich rightly notes, Kant's problems begin and end with assigning the two elements of moral insight to two separate transcendental faculties. Qua sensible creature, an agent experiences the necessary exclusion of inclination from moral choice as an external constraint. Qua rational being, the agent experiences this limitation on inclination as elevating, but only at the expense of inclination. The problem is that the agent, as the situs of these contrasting elements, does not experience its *self* as anything but an aggregate of these dissimilar elements; they are not fundamentally related to one another in a unified moral character (UR 110).

This deficit plays itself out in Kant's moral theory in a number of areas. Although Henrich points to problems deducing duties to oneself from the categorical imperative as evidencing Kant's inability to meet the demand of ordinary moral insight, Kant's idealist critics focused on a coordinate failure involving duties toward others. Even if such duties can be derived from the categorical imperative (on the first of its formulations)—that is, are universalizable—they violate the "intention of moral consciousness" (UR 102), because the "[o]rigin of duty in the limitation of our inclination has the consequence that the intention of the moral will is directed to duty by excluding the required content" (UR 102). The "required content" here is feeling for the well-being of *that* person. Now, Kant certainly does not hold that acting from duty requires the agent not to act for the sake of all sorts of praiseworthy things, including love for the recipient of the moral action; and Kantian moral agents do not need to take the simple doing of one's duty as a moral end. Duty is the required motivation, not the required end, for an agent. But the point remains that Kant thinks that the "required content" of the moral motivation can be neither positive nor negative inclination concerning the object of moral judgment. Hence the deep dissatisfaction among Kant's idealist critics with Kant's views on autonomy; it is not that idealism abandoned the idea of autonomy entirely, but that idea will have to shed its Kantian chrysalis in order to provide a satisfactory account of moral insight.

Henrich emphasizes two lines of development. In one, Friedrich Schiller argues for an intimate relation of sensible feeling to reason. Whereas Kant would only go so far as to claim that the practice of doing one's duty could result in a disposition to do good (and even that we have an imperfect duty to develop such a disposition), Schiller states that inclination *itself* can have moral content; that is, an inclination *to duty* is possible (UR 116). This response leaves intact the Kantian idea that there are transcendental faculties, but argues against the view that the faculties are ultimately in conflict. G. W. F. Hegel takes a more direct approach, construing conflict of our sensuous and moral natures on the model of unity in difference [illegible] discarding altogether Kant's reliance upon discrete cognitive faculties (UR 120-21). Kant's "moral standpoint" is but an expression of an alienated bourgeoisie that internalizes the public arena of ethics proper and, in doing so, creates an ultimately unreal private sphere of virtue.

Problems with Kant's faculty psychology also figure prominently in "The Unity of Subjectivity," Henrich's review of the second edition of Martin Heidegger's *Kant and the Problem of Metaphysics*.[7] Heidegger's Kant is a product of Heidegger's own interpretive regimen, which takes it to be its task to unearth the "unsaid" in a particular mode of thought. What is "unsaid" is equivalent to that which is true yet inexpressible within the philosophical view under consideration. An explanation of why and how the truth escapes a relevant mode of thought will involve an analysis of those systematic features of that thought which suppress truth. It is no surprise that the truth to be "said" is Heidegger's. So, the Heideggerian reconstruction of Kant will proceed along the following lines: (1) assume the correctness of *Being and Time*[8]; (2) reinterpret the transcendental subject as *Dasein*; and (3) determine what form of Kant's theory would take, given that substitution.

Heidegger's interpretation focuses on the role of the transcendental imagination in Kant's account of synthesis. That role is to produce schemata for a priori concepts that could not otherwise be displayed in intuition—a prerequisite for the deployment of a concept to a manifold of intuition for Kant. In order to effect this conjunction of pure concept and sensible manifold, the imagination has to anticipate experience "in both directions," so to speak. In order that it do so, the imagination must share characteristics of both sensibility and the understanding. Kant mentions the possibility of a "common, but to us unknown root" of sensibility and understanding in a passage in the introduction to the first edition of the *Critique of*

---

[7] *Kant und das Problem der Metaphysik* (Frankfurt a.M.: Klostermann, 1951, 2nd ed.); English translation by Richard Taft of the 4th enlarged edition published as *Kant and the Problem of Metaphysics* (Bloomington: Indiana UP, 1990).

[8] *Sein und Zeit* (Tübingen: Niemeyer, 1986, 16th ed.); English translation by John Macquarrie and Edward Robinson of the 7th edition published as *Being and Time* (New York: Harper, 1962).

Pure Reason, whose context rather strongly indicates that he actually views the root as not only unknown, but also as unknowable (A15/B29; compare A141/B180; A649/B677). Enemies of faculty psychology, such as Heidegger (and Hegel), read the root to be the imagination, on the basis of the imagination's special status as the "third man" in synthesis, and argue that it is the imagination that provides unity to the Kantian cognitive subject. Henrich shows that Heidegger can make his version of this claim only under the Heideggerian presupposition that a fundamental structure can consist of a plurality of elements that are equiprimordial (*gleichursprünglich*), that is, interdependent but nonreducible.[9] As against the idealist attempt to discover a single, *simple* faculty that was such a root, Heidegger treats the common root as a troika of imagination, understanding, and sensibility. Each faculty depends on being related on the others, which relation is made possible by the imagination. This completely alters Kant's understanding of what a subject is and, therefore, Henrich argues, we must reject Heidegger's interpretation. Kant not only has no operative conception of equiprimordiality; he could not have had one. Of course, the Heideggerian may just grant all of this—that the truth for Kant was unsaid and unsayable—without giving up the "truth" of Heidegger's Kant "interpretation." Characteristically, Heidegger never claims that he is giving, or cares to give, a "correct" rendition of Kant. It might be argued, to the extent that Kant's views cannot accommodate equiprimordiality, that the insertion of that idea has made clear (at least in part) how much "untruth" there is in Kant. If one takes this tack, Henrich's criticism does not so much amount to a demonstration that Heidegger's views on Kant must be rejected, as much as it places in clear view the substantial difference between Henrich's method of argumentative reconstruction, which still counts fidelity to the text as important, and Heidegger's interpretive method, which does not.

II

Although Wilhelm Dilthey, Heidegger, and others have stressed the philosophical importance of Hölderlin, there has been a tendency to view the history of classical German philosophy chiefly in terms of the sequential construction of the philosophical systems of J. G. Fichte, F. W. J. Schelling, and Hegel, with little place for an unsystematic thinker of self-avowed amateur status such as Hölderlin. True, Hölderlin was known to be at the epicenter of the development of German idealist thought through 1798, but his own philosophical importance was thought to be limited. This understanding changed radically upon the publication in 1961 of a writing dated 1795, entitled "Judgment and Being." Henrich's "Hölderlin on Judgment and Being" is the classic analysis of this very important fragment. Hölderlin is shown to develop a critique of Fichte which is profoundly original.

---

[9] *Sein und Zeit*, p. 131.

Hölderlin construes judgment as primordial separation (*Urteilung*) into subject and object. Prior to this separation is undifferentiated Being. Since knowledge or object-consciousness requires the opposition of subject and object, we can have no knowledge or consciousness of Being. All that can be done philosophically is to designate this unity between subject and object by a boundary concept (CR 75-76). Hölderlin's views are a mixture of Plato, Baruch Spinoza, Kant, and Fichte. The Spinozistic element is clear in Hölderlin's notion of Being as a undifferentiated totality. This element is crucial in that it places restrictions upon specific forms that an idealist theory can take. For instance, Fichte's doctrine of the posits, which explicates the preconditions for consciousness in terms of the logical relations of the I, not-I, and the mutual limitation of the I and not-I is compromised, since 'I' is a designation that relies upon differentiation. Similarly, identity theories, such as those developed by Schelling, doubly fall prey to the objection, in that identity is a relational *concept* and one requiring *distinct* relata. Yet, even given the anti-Fichtean vein, Hölderlin agrees with the Fichtean substitution of opposition for Kant's synthesis of the manifold as the fundamental mental activity. But, in spite of the critique of Kant, the passage also has a Kantian ring: Being is an indemonstrable presupposition, albeit a necessary one—not precisely what Kant would have called a "regulative, yet transcendentally necessary, idea," but close (CR 86). Last, Hölderlin's analysis of opposition and unity substitutes for Fichte's "consciousness" a Platonic conception of love as the dialectic of self-sacrifice and possession (CR 130-31). While an atavistic recapturing of original unity is impossible, the hope for reunification is pervasive and expressed in the experience of beauty. As Henrich writes:

> Man comes forth from a unitary ground to which he remains connected in the certainty of the presuppositions of his existence and of the possibility of a new unity. At the same time, he is bound to a world that, like himself also originates in opposition. For the sake of unity, he strives active beyond each of its boundaries. Yet in it he at once confronts the beautiful—an anticipation of the unity that is lost to him and that he seeks to restore. As he embraces the beautiful, the complete truth,

> which lies at an infinite distance, is realized for him within limits. He is thus captivated by it, and for good reason. But he must not forget that his active nature is called upon to overcome the finite. In the conflict of love and selfhood he runs his course, either errantly or with self-understanding (CR 127-28).

Later, Hölderlin presents a variation of this picture which reserves a special place for artistic beauty and expression. He redescribes life as having a triadic structure: (1) the self strives to perfection ("infinity"); (2) the self is limited to the finite that cannot grasp the original totality of Being; (3) aware of the necessary tension of (1) and (2), the self accepts constant oscillation between the elements. Henrich aptly characterizes this as a state in which the "tendencies of life" are in continual "modulation" (CR 134). Art recollects this process of conflict, ordering so as to confer upon it the only completeness it can have—a dynamic, fleeting intimation of unity. The artwork embodies in its pure form what Hölderlin believes to be the fundamental structure of all consciousness: its *lived* quality (CR 222). Here, Hölderlin implicitly criticizes the notion that primary access to the phenomenon of consciousness is reflective. Our true understanding of consciousness comes only from being conscious in the way that we are. Being conscious reveals a felt or intuitive grasp of the harmony of life's tendencies without attempting to forge their unity conceptually. This is the core of Hölderlin's conception of "recollection" that Henrich reconstructs in his close reading of the poem "Andenken."

Hölderlin's texts often seem as fragile as its author,[10] alembics which are themselves apt to dissolve under the pressure of philosophical reconstruction. I suspect that Henrich has excavated and developed these fragments as thoroughly as anyone should hope. In giving such a nuanced and complete account of the philosophical underpinnings of Hölderlin's thought, Henrich has staked out important ground as against rival interpretations. Heidegger views the real philosophical content of Hölderlin to reside in his late hymns and odes and construes him as essentially apart from the intellectual current of his time. Alternatively, Theodor Adorno views Hölderlin's importance to lie in his "paratactic," fragmentary style, and does not take him to be a thinker seriously concerned to forward a comprehensive account of experience. Among the many merits of Henrich's account is that it situates Hölderlin in the best philosophical company, concerned to answer what were the pressing problems of the time in a philosophical way. Further, Henrich's reading of Hölderlin adumbrates an alternative view of speculative thought sufficiently distanced from the systematic reflections of Fichte, Schelling, and Hegel which may be capable of further development. Such speculation would emphasize Kantian elements in Hölderlin, thereby treating the unity of subject (or of reason) as an indemonstrable and regulative idea, but one that we are constitutionally constrained to entertain. Speculation in this sense would be an incompletable task, yet one we must perform in order to articulate the ineliminable duality that is constutitive of our understanding of life. On this view, reconciliation is not a matter, as Hegel thought, of gaining a comprehensive enough self-understanding through which all apparently fundamental differences are found out not to be fundamental at all. Rather, reconciliation consists in accepting "the conflict of life's tendencies" as basic to our being.[11]

---

[10] Declared insane at the age of thirty-six and released to the care of an admiring carpenter in Tübingen, Hölderlin lived the remaining half of his life forgotten by the world.

Henrich's German is rendered with clarity and precision, and special mention should be made of Taylor Carman's and Jeffrey Edwards's translations. Although in the end one may differ with Henrich's reconstruction of the deduction or question the contemporary salience of Hölderlin's thought, *The Unity of Reason* and *The Course of Remembrance* reflect a remarkable philosophical depth and sensitivity. Both books are essential reading for those interested in classical German philosophy and its development, or in the relevance of the philosophical views first articulated in that tradition to contemporary concerns.

FRED L. RUSH, JR.

University of Kansas

---

[11] In broad stroke, this is Henrich's project. See, *inter alia*, "Selbstbewußtsein und spekulatives Denken," in *Fluchtlinien* (Frankfurt a.M.: Suhrkamp, 1982), pp. 125-81; "Grund und Gang spekulativen Denkens," in Henrich and Rolf-Peter Horstmann, eds. *Metaphysik nach Kant?* (Stuttgart: Klett-Cotta, 1988), pp. 83-120; "Gedanken zur Dankbarkeit," in Reinhard Löw, ed., *Οἰκειῶσις: Eine Festschrift für Robert Spaemann* (Weinheim: Acta Humaniora, 1987), pp. 69-86.

NEW BOOKS

ABEL, REUBEN: *Man Is the Measure.* New York: Free Press, 1997. xxiv, 296 p. Paper $16.95.

ACKRILL, J.L.: *Essays on Plato and Aristotle.* New York: Oxford, 1997. ix, 239 p. Cloth $60.00.

ALFORD, C. FRED: *What Evil Means to Us.* Ithaca: Cornell, 1997. xi, 185 p. Cloth $22.50.

ALLEN, COLIN, and MARC BEKOFF: *Species of Mind: The Philosophy and Biology of Cognitive Ethology.* Cambridge: MIT, 1997. xxi, 209 p. Cloth $35.00.

ANDERSON, PAMELA SUE: *A Feminist Philosophy of Religion.* Malden, MA: Blackwell, 1997. xvi, 287 p. Cloth $54.95, paper $24.95.

ARMSTRONG, D.M.: *A World of States of Affairs.* New York: Cambridge, 1997. xiii, 285 p. Cloth $54.95, Paper $19.95.

AUDI, ROBERT: *Moral Knowledge and Ethical Character.* New York: Oxford, 1997. xi, 304 p. Cloth $49.95, paper $19.95.

BAIER, ANNETTE C.: *The Commons of the Mind.* Chicago: Open Court, 1997. ix, 110 p. Cloth $26.95, paper $15.95.

BAILEY, JAMES WOOD: *Utilitarianism, Institutions, and Justice.* New York, Oxford, 1997. xii, 203 p. Cloth $49.95.

BAUERLEIN, MARK: *The Pragmatic Mind: Explorations in the Psychology of Belief.* Durham, NC: Duke, 1997. xix, 136 p. Cloth $45.95, paper $15.95.

BEDAU, HUGO ADAM: *Making Mortal Choices: Three Exercises in Moral Casuistry.* New York: Oxford, 1997. ix, 123 p. Cloth $29.95, Paper $13.95.

BENCIVENGA, ERMANNO: *Freedom: A Dialogue.* Indianapolis: Hackett, 1997. viii, 107 p. Cloth $34.95, paper $12.95.

BEST, STEPHEN, and DOUGLAS KELLNER: *The Postmodern Turn.* New York: Guilford, 1997. xiv, 306 p. Paper $18.95.

BIESECKER, BARBARA A.: *Addressing Postmodernity: Kenneth Burke, Rhetoric, and a Theory of Social Change.* Tuscaloosa: Alabama UP, 1997. x, 123 p. Cloth $29.95.

BOGDAN, RADU J.: *Interpreting Minds.* Cambridge: MIT, 1997. xii, 304 p. Cloth $35.00.

BOLOTIN, DAVID: *An Approach to Aristotle's* Physics. Albany: SUNY, 1997. 156 p. Paper $14.95.

BRETT, ANNABEL S.: *Liberty, Right and Nature: Individual Rights in Later Scholastic Thought.* New York: Cambridge, 1996. xii, 254 p. Cloth $59.95

BRUNS, GERALD L.: *Maurice Blanchot: The Refusal of Philosophy.* Baltimore: Johns Hopkins, 1997. xxxii, 339 p. Cloth $39.95.

BURGESS, JOHN P., and GIDEON ROSEN: *A Subject with No Object: Strategies for Nominalistic Interpretation of Mathematics.* New York: Oxford, 1997. x, 259 p. Cloth $45.00.

BURLEIGH, MICHAEL: *Ethics and Extermination: Reflections on Nazi Genocide.* New York: Cambridge, 1997. xi, 261 p. Cloth $54.95, paper $17.95.

BUTLER, JUDITH: *The Psychic Life of Power: Theories in Subjection.* Stanford: University Press, 1997. 218 p. Cloth $39.50, Paper $14.95

THE JOURNAL OF PHILOSOPHY

VOLUME XCV, NO. 4, APRIL 1998

LANGUAGE CONVENTIONS MADE SIMPLE

At the start of *Convention: A Philosophical Study*, David Lewis[1] says "It is a platitude that language is ruled by convention," and he proposes to give us "an analysis of convention in its full generality, including tacit convention not created by agreement" (*ibid.*, p. 1). Almost no clause of Lewis's analysis has withstood the barrage of counterexamples over the years,[2] however, and a glance at the big dictionary suggests why, for there are a dozen different senses listed there. Left unfettered, convention wanders freely from conventional wisdom through conventional medicine, conventions of art and "conventions of morality" to conventions of bidding in bridge.[3] Surely, it is unwise to try to fell these all with a single stone. Lewis's original goal, however, pursued further in "Languages and Language,"[4] was to describe the conventionality of language, and this may be a more reasonable target.

---

[1] Cambridge: Harvard, 1969.

[2] For a good discussion of the various difficulties, see Margaret Gilbert, *On Social Facts* (New York: Routledge, 1989; reprint, Princeton: University Press, 1992).

[3] *Webster's Third International Dictionary*. Senses 3, 4, and 5 alone—those most pertinent to Lewis's aims—include the following variety: "3. Agreement or an agreement; specif., an agreement enforceable in law; a contract; covenant; 4. General agreement or concurrence, as the basis of any custom, opinion or the like, or as embodied in any accepted standard, method, usage or the like; hence, arbitrary or inflexible custom; fixed usage; conventionality; as slaves to convention; 5. A rule or usage based upon general agreement; a rule or practice generally adhered to; an arbitrary or inflexible rule, form, principle, etc., as in art; a conventionalism; as, the conventions of morality or of everyday life." It is not surprising that an attempt to sum up these uses under a single "analysis of convention in its full generality" should fall to counterexamples.

[4] In K. Gunderson, ed., *Language, Mind, and Knowledge* (Minneapolis: Minnesota UP, 1975); reprinted in H. Geirsson and M. Losonsky, eds., *Readings in Language and Mind* (Cambridge: Blackwell, 1996), pp. 134-55.

0022-362X/98/9504/161–80

This article is aimed at that target. It is aimed, more specifically, at describing the conventionality of natural as opposed, say, to stipulated language. I do not claim that this kind of conventionality is the only kind of conventionality there is. I do not claim there are no other senses in which even languages, non-natural ones, might be "conventional"; and I do not claim that it is definitional of all language that it has to be conventional. (Perhaps there are animal "languages," languages that are inherited rather than conventional in the sense I shall describe.)

Why should we care about the conventionality of natural language? Because understanding it is essential for understanding many aspects of pragmatics: for example, questions in speech-act theory and about how natural-language demonstratives work. I shall remark in section VII on the relation of the analysis I shall give to speech-act theory.

Lewis's analysis of conventions was very complex, involving solutions to "coordination problems" defined in a complex way, regular conformity to convention within a group, mutual knowledge of this conformity supporting its continuation on rational grounds, and so forth. I shall try to capture the conventionality of natural language in very much simpler terms, displaying its continuity with much more rudimentary conventions. I shall call this kind of conventionality *natural conventionality* and speak, correspondingly, of *natural conventions.* Natural conventions will be described as requiring neither coordinations, regular conformity, nor rational underpinnings.[5] Natural conventionality is composed of two, quite simple, related characteristics. First, natural conventions consist of patterns that are "reproduced" in a sense to be defined. Second, the fact that these patterns proliferate is due partly to weight of precedent, rather than due, for example, to their intrinsically superior capacity to perform certain functions. That is all. I shall discuss these two characteristics in turn, explaining them as they characterize, first, simple noncoordinating conventions, then simple coordination conventions, and, finally, language conventions, most though not all of which are coordinative.[6]

---

[5] I have inveighed elsewhere against the psychological reality of Gricean intentions, hence of "mutual knowledge" in Lewis's sense—*Language, Thought, and Other Biological Categories* (Cambridge: MIT, 1984), chapter 3.

[6] It is doubtful, for example, whether saying 'Damn!' when annoyed and 'Ouch!' when it hurts are coordination conventions. Expletives are frequently uttered by persons when alone, nor can one question an explicative with 'Were you talking to me?' The conventions of saying 'boo!' to scare someone and of calling cats with 'kittykittykitty' (United States) or 'pusspusspuss' (England) or 'tsttsttst' (Hungary, Poland) are not coordination conventions. Lewis says that guiding horses with 'gee' and 'haw' is a coordination convention among drivers who must drive the same horses, however, and perhaps he is right.

I. FIRST REQUIREMENT. REPRODUCTION

The various conventional aspects of natural-language forms are reproduced structures. Whole sentences and larger units are not usually reproduced, of course; but words are reproduced and—even if we humans all come with a Chomskian inherited universal grammar—many aspects of syntax are also reproduced. A pattern has been *reproduced* if its form is derived from a previous item or items having, in certain respects, the same form, such that had the model(s) been different in these respects, the copy would have differed accordingly. A reproduction is never determined by its model in all respects. The photocopy has a white background not because the original did, but because white paper was put in the copier's paper tray. Similarly, if certain features of syntax are determined by the endogenous structure of the language module, this does not prevent other aspects of syntactic patterns from being reproduced. It is by reproducing language forms that children growing up in France learn to speak French, children in China, Chinese.

Being perpetuated by reproduction is the first basic feature of natural conventionality. Natural conventions are handed down. Consider, by contrast, the *shirt-buttoning* way of proliferating a pattern. Suppose that we all button our shirts in order from top to bottom, but that we do so quite independently, each having independently discovered for herself that this is the easiest way to get the right buttons into the right holes. This convergence of behavior would not result from our shirt-buttoning techniques being handed down, hence would not be conventional behavior. Behaviors that are idiosyncratic or peculiar seem quite certain not to have been handed down, hence these are the paradigm nonconventional behaviors—the *unconventional* behaviors.

On the river in Cambridge, England, it is conventional to punt standing on the deck at the back of the punt boat whereas in Oxford, one turns the deck to the front and stands at the back inside. In both places, you see many other, more awkward, antics besides, performed both by undergraduates and visiting untutored novices. Most untutored novices, however, either stand on the deck in back or turn the deck in front and stand behind and inside. In so doing, they are not *following* the tradition handed down by either university. Rather, their behaviors accidentally coincide with these traditions. Their behaviors are no more conventional than the behavior of a puppy who accidentally wets in the bathroom. A behavior is conventional not because its form matches a conventional one but because its form was produced by reproduction.

There are a number of different ways in which conventional patterns can be reproduced which will be of interest to us. They may be reproduced by being copied from one another directly, or one person may tell another how a pattern goes. For example, Johnny's mother tells him that he is to put his letter in the mailbox and put up the flag. The result is that he reproduces the first part of a conventional pattern of activity, the second part of which will be reproduced by the mailman, who, on seeing the flag, stops at the box, removes its contents, then puts the flag down. He does this, in part, because he is familiar with this conventional pattern of activity and, recognizing that the first part has been performed, reproduces the second part. With his addition, the full conventional pattern has now been reproduced.

Sometimes conventional patterns are reproduced by what I shall call *counterpart reproduction*. A bolt one wishes to be suitable to a given purpose usually has various thread gauges as reasonably functional alternatives, but once a particular nut has been chosen, the choice of suitable thread gauges is strictly determined. Likewise for the nut. So if certain nuts were made to fit certain bolts, and then more bolts made to fit the nuts, and so on, specific thread gauges of both nuts and bolts would be reproduced without any bolt's being directly copied from any bolt or any nut from any nut. The traditional positions assumed by men and women for ballroom dancing were commonly reproduced in part by *nuts-and-bolts reproduction*, each woman settling into the traditional woman's posture in response to the postures of the men with whom she danced, and vice versa. A simpler kind of counterpart reproduction is *handshake reproduction*. Shaking hands by placing the right hand in another's right hand, thumb notch to thumb notch, then shaking up and down (contrast left hands, grasping wrists, interlocked fingers, shaking side to side) is probably learned in part not by direct copying but by fitting in with one's partner; but here the counterparts must be alike to fit. Often each individual part needs to fit arbitrary members from a whole collection of counterparts, and vice versa. This easily results in standardization of forms, more easily than by direct copying. Copies of copies easily drift away from the original; the need to fit counterparts retards drift.

That a certain pattern of behavior is common in one culture but not in another is evidence, but not particularly strong evidence, that the pattern may be being proliferated by reproduction. One reason it is not particularly strong evidence is that the only practical manner of doing a thing may depend upon context and hence vary from culture to culture, proliferated in the shirt-buttoning way. The sensible way to dress in Iceland is not sensible in Ecuador, copying quite to one side. But the reasonable way to behave may also depend upon prior patterns already established in one's culture, so that a pattern learned merely as practical, as a skill, at a deeper level of analysis may be reproduced by counterpart reproduction. Margaret Gilbert says that in some circles it is conventional to write a thank-you note after attending a dinner party but that, contra Lewis, there is nothing arbitrary in doing so (*op. cit.*, p. 340). Indeed, if the host expects you to write such a note and will think you rude or ungrateful if you do not, then you have reason to write it, conventions quite to one side. On the other hand, the host will have such an attitude only because people before you have written thank-you notes under such circumstances. Had people before you made phone calls, or sent messages or flowers, your host would have expected that instead. Thus, even if you were not aware of doing what others have done, writing a note in response to your host's expectations, however you chanced to discern them, is a reproduced behavior; it is handed down. Reproduced patterns are often learned rather in the manner of skills, or because obviously morally required, and so forth, while at a deeper level they are portions of counterpart-reproduced, conventional patterns.

Patterns of activity are often reproduced unconsciously. For example, social distances—the distances at which people stand when conversing—vary from culture to culture and from one kind of social relationship to another within each culture. If you are standing at the wrong social distance, the person to whom you are talking will move; so to prevent slow circling about the room as you talk, you unconsciously reproduce the conventional social distance by handshake reproduction. Shaking hands with the right hand, specifically, may sometimes be reproduced in this unconscious way. One can even imagine being forced, without realizing it, into the right-hand driving pattern merely in an effort to avoid on-coming traffic. Then one would unconsciously reproduce a driving pattern by handshake reproduction.

The reproduction and standardization of natural-language patterns may take place, in part, through unconscious counterpart reproduction. Reproduced speaker-hearer patterns generally involve inner as well as outer acts so that direct copying is not a means of reproduction for the whole. But given the speaker's portion of the pattern, the hearer's part is not arbitrary; and given the hearer's portion, the speaker's part is not arbitrary. So the ability to play either of these roles may be acquired, in part, as skills, given partners who are reproducing the complementary portions of these patterns. For example, those complex patterns whose standard job it is to transfer information—involving, paradigmatically, the use and interpretation of indicative sentences—are completed as the hearer performs the inner act of interpreting these sentences, in accordance with certain rule-like patterns, into beliefs. This movement is a hidden action and (René Descartes to the contrary) not voluntary.[7] Hearers learn to believe in conformity with the conventional rules of the language because this often brings the reward of useful knowledge gained. Except for the (probably very large) boost from certain inborn language capacities (compare the texture and color of the paper fed into the copying machine), they learn it much as they learn to interpret natural sign patterns. Similarly, speaker-hearer patterns that go with paradigmatic uses of imperatives begin when the speaker expresses a purpose with regard to the hearer's actions with outer signs in a certain established, that is, reproduced, manner. Learning to do this may result, in part, from the speaker observing the effects that his imperatives, and those of others, have on hearers who are already disposed to complete these patterns by supplying conforming actions in the established manner.

II. SECOND REQUIREMENT: WEIGHT OF PRECEDENT

Not all reproduced behaviors are conventional behaviors. Handing down a skill is not, as such, proliferating a convention. I learned from my mother, and she from hers, to open a stuck jar lid by first immersing it in hot water. Opening jars this way is not thereby "conventional." To be thought of as conventional, a reproduced pattern must be perceived as proliferated due, in important part, to weight of precedent, not to its intrinsically superior capacity to produce a desired result, or due, say, to ignorance of any alternatives.

Thus, a pattern is considered conventional only if thought to have little tendency to emerge or reemerge in the absence of precedent. This might be because the pattern has no useful function at all: wearing black and white to examinations at Oxford University, decorating for Christmas specifically with red and green, handing out, specifically, cigars after the birth of a boy. Or this might be because there are other viable patterns of behavior that could have served the same functions equally well had precedent for them been established. Thus, a certain arbitrariness of pattern in relation to function—evidence for this perhaps gleaned from other cultures, where other patterns do serve just as well—is an argument for conventionality. Conventional patterns are patterns for which other patterns, given different historical accidents, might as well have been substituted. Lewis, too, emphasizes arbitrariness as essential to conventionality.

---

[7] Compare Lewis's claim that forming a belief is not a voluntary action and hence cannot be part of any convention (*Convention*, p. 180). Later ("Languages and Language"), he allows beliefs formed on rational grounds to be part of the "content" of language conventions. But social distances are not adopted in accordance with reasons. Indeed, they are so unconscious that their very existence is known to few people and went unremarked until anthropologists began to study them in the 1950s, yet they are surely paradigms of natural convention.

Yet an arbitrariness of pattern in relation to function is not sufficient evidence for conventionality. The songs of the various bird species are to a large degree arbitrary in relation to function, but they are not conventional because they are not copied or reproduced in the sense defined above. That other reproduced patterns would serve just as well if only their possibility was *known* is also not an argument for conventionality. Lighting fires by rubbing two sticks together may be a sort of convention among boy scouts, but it is not merely conventional in a culture that knows no other way to light fires. Similarly, conventional Western medicine is called *conventional* only because other kinds of medicine are now also known in the West. Calling it conventional also implies (perhaps euphemistically) that these other kinds are genuine alternatives, only weight of tradition serving to exclude them from more common use. But where knowledge of alternatives is unavailable, precedent figures in the proliferation of a pattern not essentially, but only as a means of making its possibility known. If people discovered the pattern's potential independently of one another, it would proliferate just as readily. A conventional pattern, by contrast, proliferates partly due to precedent.

There are various mechanisms through which patterns of activity may proliferate due to weight of precedent. A pattern may prevail over easily invented alternatives merely because it is easier or more natural to copy than to use one's imagination, or because people prefer to do as others do, not wanting to be out of step, or because what is familiar is as such pleasing, or because people feel more secure in the tried and true. In some cultures, all traditional behaviors may be positively sanctioned as such, either out of simple respect for tradition, or for fear of disturbing the moral order. It can also be habitual to place sanctions on certain but not other conventional behaviors, for approval and disapproval themselves are often copied without much thought. In all such cases, weight of precedent itself is helping to account for the proliferation of the pattern, making the pattern conventional. Sanctions rooted merely in tradition often are not recognized as such but are thought to be independently based in morality, or etiquette, good taste, necessity for health, decorousness, propriety, piety, and so forth. Often there is room for reasonable disagreement over which reproduced behavioral patterns are sanctioned merely conventionally and which are sanctioned due to some function they are serving.

When patterns of behavior are proliferated by counterpart reproduction, often it is weight of precedent which keeps them afloat, and which does so for very practical reasons. In the Orient, because people have learned to use chopsticks, they are the implements placed on the table for dining, manufactured in quantity, available easily and cheaply, and so forth. In part because they are manufactured in quantity, available easily and cheaply, and so forth, chopsticks are placed on the table and people learn how to use them. Similarly for forks in the West. Weight of precedent proliferates their use, and not just because people are conformists. Proliferation of the use of standard units of measure is another example of this sort. Measurements taken in conventional units proliferate in tandem with the calibrated devices used to make the measurements. For example, measurements taken in yards proliferate as yardsticks proliferate, and vice versa. Weight of precedent proliferates the use of conventional units of measure also by handshake reproduction. Measurements taken in yards are easiest to compare or to calculate with other measurements taken in yards, those taken in meters with others taken in meters, and so on.

III. COORDINATION CONVENTIONS

Lewis claimed that all conventions solve coordination problems among the participants.[8] A subclass of natural conventions, I believe, are indeed *coordination conventions*. My description of coordinations will be simpler than Lewis's, however. Especially, no beliefs on the part of the participants will figure as essential ingredients in the analysis.

There is a need for coordination when:

(1) members of a group ("partners" in a projected coordination) have a purpose in common;
(2) achieving this purpose requires actions by each of the partners;
(3) more than one combination of actions will achieve the purpose;

---

[8] *Convention*, pp. 36ff.

(4) the set of workable combinations fails fully to determine what any single partner's contribution must be independently of the actions performed by the others.

Coordination is achieved if the partners' combined actions achieve the common purpose. Coordination conventions are conventional patterns of activity that proliferate, in part, (causally) because they achieve coordinations.

Some coordinations require the partners to act the same way, others to act differently. Coordinating social distance requires acting similarly. Although each person learns independently to maintain the conventional distances (as we may learn independently to button our shirts from top to bottom), it is not so that keeping just these distances has intrinsic superiority over keeping other distances, nor that ignorance of other possibilities is a factor in proliferating the distances actually kept. Rather, that others are already doing things this way causes new participants to follow after. Thus, mere weight of precedent perpetuates this coordination solution. Another coordination, discussed by Lewis, which requires similar actions, is a community's using certain standard media (minted money) for exchange of goods. I have mentioned that common units of measure are proliferated in part by handshake reproduction because of the ease of comparing measurements made with a common measure (the "partners" here might be oneself at different times). As such, the use of standard units of measure is a conventional coordination. In another case discussed by Lewis, different actions are required by the partners to achieve coordination. When a telephone connection is broken, one person must call back while the other waits.

On the other hand, we could say of the telephone partners that they must do the same thing, namely, follow the pattern: original caller calls back, original receiver waits; or perhaps original receiver calls back, original caller waits; or perhaps, whoever's sentence was interrupted calls back, the other waits. If any one such pattern should begin to appear slightly more frequently than others, due perhaps to the vagaries of statistics with small numbers (compare "evolutionary drift"), granted this is consciously or unconsciously noticed, it will tend to reproduce in a counterparts manner because people learn from experience. Then, like a gene with a fitness advantage, it may "tend to fixation." That others are already doing things this way causes new partners to follow after.

Gilbert has argued that there is nothing rational about following precedent in the attempt to solve a coordination problem, unless one makes more assumptions about the nature of one's partners than just that they are rational.[9] What more one needs, I am suggesting, is to assume that one's partners routinely adapt themselves to their environment by repeating what has worked in the past. Coordination conventions proliferate because, rationality aside, people learn from experience exactly as other animals do.

IV. CONVENTIONS DO NOT IMPLY REGULAR CONFORMITY

There is a long tradition of taking conventions *as such* to involve regular conformity within a group either de facto or de jure.[10] If conventions are as I have described them, this is a mistake. In this section, I shall argue against regularities de facto; in section V, I shall argue against regularities de jure.

That people need not regularly conform to noncoordinating conventions is clear. Few actually hand out cigars at the birth of a boy, nor does everyone wear green on St. Patrick's Day, or decorate with red and green on Christmas, or punt from the deck when on the Cam River. What, then, of coordinating conventions?

First, I shall argue that to solve a coordination problem it is not usually needed that all sets of cooperating partners in some group solve it in the same way. It will help here to differentiate among three different kinds of coordinations that can be proliferated by weight of precedent. I shall call these *blind coordinations*, *half-blind coordinations*, and *open coordinations*. In blind coordinations, each partner must act before having any evidence concerning the other's action(s). Thus, when a telephone connection is interrupted, neither party has information about the other's reaction until the coordination has either succeeded or failed. Contrast two people moving a couch together where each sees the anticipatory motions of the other. Usually, one adopts a leader role, the other a follower role, and an excellent open coordination may be achieved in the absence of any conventions at all. Between these extremes is half-blind coordination: for example, when a child raises her hand to speak in a classroom and is called on by her teacher. Here, part of the conventional pattern wishing-to-speak-produces-hand-raising-produces-being-called-on which the child initiates is visible to the teacher, but the wishing-to-speak part is not. Only conventions for achieving the first kind, the totally blind coordinations, depend for effectiveness on a degree of regular conformity within an interacting group.

---

[9] *On Social Facts*, chapter 6; "Agreements, Conventions, and Language," *Synthese*, LIV (1983): 375-407.

[10] John Searle, *Speech Acts: An Essay in the Philosophy of Language* (New York: Cambridge, 1969); Lewis, *Convention* and "Languages and Language"; Stephen Schiffer, *Meaning* (New York: Oxford, 1972); Kent Bach and R. M. Harnish, *Linguistic Communication and Speech Acts* (Cambridge: MIT, 1979); Gilbert, "Agreements, Conventions, and Language" and *On Social Facts*; François Recanati, *Meaning and Force: The Pragmatics of Performative Utterances* (New York: Cambridge, 1987). Recanati takes it that conventional language devices "indicate" or, using the linguists' term, "mark" uses of language; that is, conventions mandate that these devices shall be used only for those purposes. This is not explicitly stated, but see, for example, §22.

Consider the right- (or left-) hand driving convention. The cooperative purpose requires that drivers approaching one another both either drive on the right or drive on the left. There is no secure way for leader-follower roles to be established, the time for decision is small, and it is hugely important that the coordination be achieved. The coordination thus behaves like a fully blind one. The only way to achieve it with complete reliability for every pair is to achieve it always the same way within some group that can easily identify its own members (people driving in America; people driving in England). Regular conformity to the same pattern is required among members of the entire larger group from which the various cooperating pairs emerge.

Contrast with blind coordinations what I shall call *leader-follower coordinations*. Here, there is no need for regular conformity. One member leads off in a manner either wholly or partially observable to the other, the other responds so as to reach a mutually desired conclusion. Many such coordinations are completely open, hence completely unproblematic, as when you seat yourself at an arbitrary table in the restaurant and I follow after. The interesting ones, those which are helped by conventions, are those which are partially blind. Consider the couch-moving case more carefully. The follower cannot be guided by the very actions with which he is simultaneously coordinating, but only by their natural precursors. It is natural, then, that the leader should exaggerate and stereotype his anticipatory movements so as to make them easier to follow. Exaggerated and stereotyped anticipatory movements, reproduced by experienced cooperators because they have been previously experienced to produce effective coordinations, are conventional*ized* movements.[11] One more step gives us conventions proper. Let the leader's indication of the action pattern he purposes for himself and his follower be divorced from any natural anticipatory movements. That is, let the reproduced pattern—it begins with a lead from a leader, proceeds through a specified pattern of action involving roles for both leader and follower, and culminates in a certain result—have an arbitrary beginning. Then it is a fully conventional coordination pattern. But there is no need for any whole group regularly to use this pattern for it to be effective in individual cases.

---

[11] The ethologist Niko Tinbergen considers the communicative signals of animals to be "ritualized" versions of "intention movements"—"'Derived' Activities, Their Causation, Biological Significance, and Emancipation during Evolution," *Quarterly Review of Biology*, XXVII (1952): 1-32. An *intention movement* is one from which it is possible to infer what an organism is likely to do next. *Ritualization* is a process whereby specific features of such movements become progressively exaggerated. It is considered likely to occur when the benefits to an organism of conveying information about its future behavior exceeds the cost of doing so. For instance, conveying information about an intention to fight may be beneficial if it prevents the fight from occurring.

Conventional leader-follower coordinations begin when a leader reproduces a certain portion of a traditional pattern, which portion is observable to a follower. The follower is familiar with the pattern, recognizes it, and reproduces the complementary part, resulting in a coordination of a sort that is partly responsible for the proliferation (due to precedent) of the pattern. When the convention is half-blind, the leader may also have completed an important part of the pattern beforehand but it was unobserved by the follower. Thus, the leader of the mailbox-flag convention places mail in the mailbox before the postman arrives, and the leader of a conventional-language pattern has invisibly made a conventional transition from belief or intention into an outer language. Or perhaps a portion of the leader's part is completed only later on, as when a promise is made.

To achieve such a coordination, the leader and follower must, of course, be reproducing parts of the *same* conventional pattern. But there is no need for only one pattern to be used in one group. There are an indefinite number of ways to impart any given piece of information using English. It is not even necessary always to speak the same language, even with the same partner. Leader-follower conventions emerge where coordinations are half-blind, but also half-sighted. Enough is seen for the follower to discern with which conventional pattern he is dealing immediately on observing the lead. Following is like chiming in after the first line of a song; there is no need for a community to know just one song. Nor does this sort of convention require that there be some definite group in which it is mutually believed that it is a convention. People acquainted with a certain leader-follower coordinating convention need not form a social group, or always know who one another are. It is usually best to know in advance whether I shall be understood by my follower, but the brute existence of a convention does not require this in each case.

Perhaps it will be thought that I have missed the essential regularity that is required by leader-follower conventional coordinations, namely, regular follow through by both parties in completing conventional patterns once initiated.

Again, consider the case of blind coordinations first. Even there, only regular *enough* observance is required to sustain them. If only one in five friends follows the telephone convention, I shall still be somewhat better off following it myself. Similarly, even if only one in fifty folks that I ask 'Give me a dime, please, just a dime?' actually gives me a dime, it may be quite sane to keep at it. On the other hand, if more than a certain critical mass of statements were (nonobvious) lies, our indicative mood forms might indeed die out. I shall say more about this below.

V. CONVENTIONS WITHOUT PRESCRIPTIVE RULES

Complementing the claim that conventions require regular de facto conformity among members of a group is the common claim that conventions rest on rules requiring conformity de jure.[12] In what sense then do conventions involve rules, and what kind of rules are these? It will be well to proceed quite carefully here.

There are many conventional patterns that are usually reproduced by explicitly following a set of rules. This is true, for example, of the conventional patterns constituting most games and of the pattern that is parliamentary procedure. Of course, it is not just the having of rules that makes a game conventional, as is shown when you make up rules for a game, thus playing an unconventional game. Notice also that rules of this sort concern regularities only within conventional patterns, not regularities, either de facto or de jure, in their being used. The rules of chess do not tell you when to play chess, nor do the rules of parliamentary procedure tell you when you must use it. It is just that you are not reproducing that particular conventional form unless you produce a pattern showing that internal regularity: in chess, the king always moves just one square; in parliamentary procedure, the chair never speaks to a motion; and so forth. On the other hand, some conventions are indeed described by rules telling when they are to be applied. There is the conventional rule, "on Christmas, decorate with red and green," and the rule, "when you are the new father of a boy, hand out cigars." In neither kind of case, however, are conventional rules, just as such, prescriptive. These rules describe conventional patterns; they do not prescribe them. The bare existence of a convention neither mandates, nor gives permission for anything.

On the other hand, conformity to a convention may be or become mandated and sanctioned. Following parliamentary procedure is mandated by law in the United States Congress. Crucially important blind conventional patterns, like driving on the right, are heavily sanctioned (indeed, often written into law), and it is not a mere matter of convention that they are sanctioned. Whenever it is desirable that some predictable social pattern or other be stabilized for some purpose in a society, or that some convention or other be universally followed, then it is not merely conventional that there are sanctions attached to conformity. But there are also conventions that are sanctioned merely conventionally, that is, sanctioned for no reason beyond tradition; and in some societies all or most conventional behavior may be sanctioned simply as such.[13]

---

[12] Gilbert, for example, explicitly endorses the latter claim but not the former.

Just as rules for constructing conventional patterns do not, as such, mandate that you must construct them, they do not, as such, prohibit constructing part or parts of the pattern without constructing the rest. The rules of chess do not mandate that having started a game you cannot then quit in the middle, or that you cannot set up an end game or a middle game and just play that much for the practice. If you cannot quit or change rules in the middle, that will be a matter of external rules or sanctions—say, rules of etiquette, or tournament rules, not the conventions of chess.

Conventions often involve repetitions of complex patterns. But fragments of these patterns may also occur quite freely. Conventional patterns are very frequently broken. This is especially common when construction of a conventional pattern requires the cooperation of more than one person. Consider again, 'Give me a dime, lady, just a dime?' That the pattern, if unbroken, would involve transfer of a dime is clear. The syntactic pattern for directive sentences continues to be reproduced only insofar as the whole speaker-hearer pattern in which it is conventionally embedded is sometimes completed. If nobody ever complied with another's directives, the directive forms in a language would soon cease to be used, indeed, to be understood. Earlier, I made a similar remark about hearers' believing assertives. It does not follow that complying with directives or believing assertives is, as such, mandated by a "rule of language" understood as prescriptive.

---

[13] The distinction is blurry between (1) patterns that are first merely conventional and only later written into law or otherwise mandated and (2) partially arbitrary patterns that originate only when written into law or otherwise mandated, resulting in a blurriness in what counts as conventional. Compare, for example, the pattern of behaviors that constitutes getting married in a certain state of the United States with the pattern to which one must conform to arrive at United States citizenship. Both are to a degree arbitrary. Neither is nowadays reproduced merely due to weight of precedent, but the marriage ceremony once was. Which, if either, is conventional? If either is conventional, might there be another sense of 'convention' edging in here, say, the kind that, according to *Webster's Third*, is "formed by agreement"? This sort of blurry line between natural conventions that are sanctioned and what we might call *regulated conventions* may help to produce the illusion that natural conventions are de jure.

Lewis tells us that the conventions of language are "truthfulness" and "trust,"[14] which unpacked means roughly: "Say only what is true or what you actually want the hearer to do," "Believe what you hear," and "Do what you are told." He is right, I believe, that these rules elliptically describe conventions for certain of our most basic language forms, but he is not right that these rules describe regularities, nor, I am now arguing, do they mandate conformity. Certainly, there are circumstances under which people are under mandate to do what they are told, and under some political regimes, even to believe what they are told. But this has nothing to do with the conventions of language.

VI. CRISSCROSSING CONVENTIONS

Another thing that the rules describing conventional patterns do not mandate is that these patterns or portions of them cannot be used outside the conventions. Broken fragments of conventional patterns are frequently used purposefully in secondary ways. They are literally or figuratively "played with" in ways requiring only that the participants understand from what wholes they were taken. Compare uses such as in word games, irony, metaphor, and so forth. Pieces having the same form as those used in one conventional pattern may be used also in constructions that have nothing to do with that convention. Sometimes they are portions of other unrelated conventions. Raising one's hand is conventional in order to request to speak, but it is also conventional in order to vote. That it is conventional to use the shape or sound 'bank' to refer to river banks does not prohibit its use in another convention to refer to financial institutions, nor does it prohibit its use in nonsense rhyme.

Where conventions overlap, using some of the same forms, it is not a matter of convention which if any is being instanced on a given occasion. That a token of any form instances a convention or piece of a convention is a matter of its individual history, not a matter of what it matches. The rules of chess do not tell when moving my queen next to your king constitutes playing that she kisses your king instead of putting your king in check, nor do any other conventional rules do this job. Only tokens reproduced due to weight of precedent are conventional, and which convention each instances depends on the precedent from which it is derived. Which if either of the hand-raising conventions is instanced when a particular hand goes up depends on which if any previous instances of hand raisings are the causes and models for this one—whether instances used for requesting to speak or instances used for voting. Which, if either, conventional word 'bank' is being used depends on which family of previous instances this instance was copied from, not on some meta-convention or rule.[15]

---

[14] *Convention*, pp. 148ff; "Languages and Language," pp. 137ff.

A language consists in a tangled jungle of overlapping, crisscrossing traditional patterns, reproducing themselves whole or in part for a variety of reasons, and not uncommonly getting in each other's way. Places where these patterns cross can produce ambiguities. These are sorted out not by conventions, but by the hearer managing to identify, by one means or another, the source of the pattern, that is, from which family it was reproduced. Nothing guarantees recognition of these sources. There cannot be, for example, such things as absolute semantic markers in a language, that is, forms reserved *merely by convention* to serve always the same function. No convention can prevent the possible accidental incursion of other crossover conventions. Where conventions cross, that the conventional pattern piece actually instanced is usually identifiable by the hearer is a function, mainly, of the speaker's skill, not of any conventions or rules of language. It is not conventional, for example, but surely good sense, not to raise your hand to speak when a hand vote is being counted. True, there can be conventions that ride piggyback and concern the use of other conventions. To avoid people taking back their hand votes later by pleading lack of intent, a rule might be made or another convention arise that when a hand vote is taken, no matter why a person put up his hand, it is counted. Still, as Dan Sperber and Deirdre Wilson[16] have rightly insisted, understanding a language is never just decoding.

VII. "CONVENTIONAL ACTS" AND THE SPEECH-ACT TRADITION

As I have described natural conventions, a conventional act is an act of reproducing or helping to reproduce, in whole or in part, a conventional pattern. Thus, people who reproduce conventional patterns do things that have been done before, and presumably were once done for the first time, thus done at least once without precedent. Conventions, I have claimed, really are nothing whatever beyond reproduced patterns of perfectly ordinary activity. Perfectly ordinary activities, like putting a letter in the mail box and then putting up the flag, or taking a letter out of the box after seeing a flag, are none of them such that they could not in principle, perhaps accidentally, occur in the absence of conventions. It follows that conventional acts achieve natural ends, but achieve them in conventional—that is, in reproduced—ways. I can set my table, hold my fork, and use my napkin in conventional ways or in unconventional ways. This was apparently the way H. P. Grice was thinking about acts of communication in "Meaning."[17] I was following a different but parallel route earlier when I described the possible evolution of fully conventional leader-follower coordinations from exaggerated and stereotyped natural anticipatory movements.[18] Given this sense of "conventional act," proceeding with a bridge game in the conventional manner after someone has said 'I bid two diamonds', and also believing what one is told and doing what one is told (these latter being traditionally counted among the "perlocutionary acts"), are all conventional acts in exactly the same sense as are, say, conventional biddings and tellings, traditionally labeled "illocutionary acts."

---

[15] Compare my *op. cit.*, chapter 4; and also David Kaplan, "Words," *Aristotelian Society Proceedings*, Supplementary Volume LXIV (1990): 93-119.

[16] *Relevance: Communication and Cognition* (Cambridge: Harvard, 1986).

The tradition of speech-act theory has something quite different in mind with the term 'conventional act'. John Searle,[19] for example, tells us that, unlike illocutionary acts, perlocutionary acts are never conventional. From J. L. Austin through P. F. Strawson, Searle, and Stephen Schiffer, to Kent Bach and Robert Harnish and Donald Davidson,[20] this tradition has taken it that "conventional acts" are acts the performance of which would be strictly impossible in the absence of conventional rules. Nor is it meant, merely, that you cannot do a thing in a conventional way unless there exists a conventional way to do it. Rather, there are supposed to be conventions that "constitute" certain natural acts (for example, raising your hand) as being, also, different acts that are conventional acts, such as acts of voting or requesting to speak, or of giving someone orders, or of asserting something, or of christening a ship, or of making a motion in a meeting, or of marrying someone. Thus, Bach and Harnish (*op. cit.*, p. 109), closely following Searle,[11] claim that a convention requires that there be circumstances such that it is mutually believed in some group that whenever such-and-such action is performed in those circumstances, a certain conventional outcome is the result: for example, the action is automatically "counted as" a marrying, or a christening. Strawson says of conventional speech acts, "the speaker's utterance is not only *intended* to further, or affect the course of, the practice in question in a certain conventional way; in the absence of any breach of the conventional conditions for furthering the procedure in this way, it cannot fail to do so" (*op. cit.*, p. 612). On these grounds, Strawson, Bach and Harnish deny of many illocutionary acts that they are conventional, replacing convention with Gricean intentions in the analysis of these speech acts. Davidson actually denies that any language is conventional on the grounds that its (ordinary) uses do not result in conventional acts of this sort (*op. cit.*, p. 276). I close this essay by offering a description of phenomena generally taken to be paradigms of "conventional acts" in this latter strong sense, but a description that shows them instead to be constituted merely by natural conventions of the kind I have been describing.

---

[17] *The Philosophical Review*, LXVI, 3 (1967): 377-88; reprinted in J. Rosenberg and C. Travis, eds., *Readings in the Philosophy of Language* (Englewood Cliffs, NJ: Prentice-Hall, 1971), pp. 436-43.

[18] A critique of Grice's way of doing this is in my *op. cit.*, chapter 3.

[19] *Intentionality: An Essay in the Philosophy of Mind* (New York: Cambridge, 1983), pp. 178-79.

[20] Austin, *How to Do Things with Words* (New York: Oxford, 1962); Strawson, "Intention and Convention in Speech Acts," *The Philosophical Review*, LXXIII, 4 (1964): 439-60 (reprinted in Rosenberg and Travis, pp. 599-614); Searle, *Speech Acts*; Schiffer, *op. cit.*; Bach and Harnish, *op. cit.*; and Davidson, "Communication and Convention: Moods and Performances," in *Inquiries into Truth and Interpretation* (New York: Oxford, 1984), pp. 265-80.

These seemingly extraordinary acts (voting, christening someone, giving orders), I shall argue, reflect little more than a way that we have of *classifying* some kinds of conventional, and also some kinds of regulated, patterns of social behavior. Examine any part forming the beginning, or occurring in the middle, of some conventional pattern of activity. There will be a piece or pieces of the pattern that conventionally follow after this part. Or perhaps this part merely puts constraints on what can follow while conforming to the convention. Call the chosen part a *conventional move*, and call what must follow its *conventional outcome*. A move while playing chess or solitaire is a conventional move in this sense. It places restrictions on what can, conventionally, happen after. Telling someone to do something is a conventional move. Its conventional outcome is that that person complies. Saying 'the meeting is adjourned' in the right circumstances is a conventional move. Its conventional outcome involves, for example, that no more motions are made and no more votes are taken.

Besides conventional patterns of activity, there are other patterns that are regulated by law or other authority: for example, the patterns of activity that constitute applying for citizenship or for a visa and its being granted or not granted, and the patterns involved in entering into a legal contract along with what legally follows. Call any parts of such patterns that have mandated outcomes or outcomes that are constrained by authority *regulated moves*, and call the mandated or constrained outcomes *regulated outcomes*. There are also many moves that are hybrid between being conventional and being regulated, a good example being traditional marriage ceremonies. Certain kinds of conventional move outcomes, and likewise many hybrid and regulated outcomes, consist in the honoring by the society at large or its representatives of certain rights and privileges, or the carrying out of certain acknowledged duties, that is, more generally, in the imposition of certain restrictions on the behaviors of people designated as involved.

---

[11] *Speech Acts*, pp. 51-52.

Now, often there are numerous different existent conventional or regulated patterns, reproduced or produced in different cultures or perhaps in the same culture, which contain differently constituted initial or middle portions, but which have the same or similar conventional or regulated outcomes. Then we are likely to possess verbal tools for classifying these conventional or regulated moves according to outcome rather than structure. Rather than talk of the natural actions that are initial or middle parts of these conventional or regulated patterns, such as raising one's hand or uttering certain sounds, we speak of voting (how?), or saying that it is raining (in what language?), or ordering someone to leave, or performing a marriage ceremony, or paying back a debt, or appointing a committee, and so forth.

In such cases, because the actions are described or identified only by reference to their conventional outcomes, of course they must be actions that are "constituted as what they are"—that is, correctly placed under these labels—due to the existence of certain conventions or regulations. They are, as such, *essentially conventional* (or *essentially regulated*). Still, there is nothing more here than perfectly ordinary natural convention and perfectly ordinary social regulation—nothing, that is, except terminology. These acts are still just ordinary actions that have been reproduced after a pattern, or following a regulation, in no way any more or differently conventional than is the act of shaking hands with the right hand, or believing it when you are told it is raining. Throwing this sort of terminology into the passive, we say that people have been appointed to committees, or are married, that the meeting is now adjourned, and so forth. That is, conventional patterns have been initiated that have certain sorts of conventional outcomes. (That these patterns have these conventional outcomes does not imply, of course, that they always or even usually come out that way.)

These oversimplified remarks may make some start, though rather a small one, toward understanding what speech acts are, and what their relations are to language conventions. Elsewhere,[22] I have argued that the central criterion that we use in categorizing speech acts does not, in the first instance, concern the conventional outcome of the act but rather the purpose of the words uttered. This purpose often coincides, of course, with the conventional outcome. The purpose of a spoken set of words, however, has two independent sources: (1) the intention of the speaker; and (2) the purpose or proper function, in the public language, of the expressions used. In central cases, both of these purposes will coincide with the conventional outcome of use of these expressions, but any two or all three of these can also come apart.

RUTH GARRETT MILLIKAN

University of Connecticut

---

[22] See my "Proper Function and Convention in Speech Acts," in L. E. Hahn, ed., *P. F. Strawson: The Library of Living Philosophers* (Chicago: Open Court, forthcoming), and my "Pushmi-pullyu Representations," in J. Tomberlin, ed., *Philosophical Perspectives*, Volume IX (Atascadero, CA: Ridgeview, 1996), pp. 185-200 (also in L. May and M. Friedman, eds., *Mind and Morals* (Cambridge: MIT, 1996), pp. 145-61).

HOW GREAT A GOOD IS VIRTUE?*

Claims about the good, and especially about the human good, are standardly divided into two classes: *subjective* or *welfarist claims*, and *objective* or *perfectionist claims*. Welfarist claims make each person's good depend on certain of her subjective mental states, such as her pleasures or desires. Hedonism, which holds that only pleasure is intrinsically good, is a version of welfarism, as are views that equate a person's good with the fulfilment of her desires. Perfectionist claims, by contrast, hold that certain states of humans are good objectively, or independently of their connection to pleasures or desires. Thus, perfectionists have held that knowledge, achievement, and deep personal relations are intrinsically good regardless of how much a person wants or enjoys them, and that their absence impoverishes a life even if it is not a source of regret.

A common perfectionist claim is that moral virtue is intrinsically good and moral vice intrinsically evil. Benevolence, courage, and similar traits in themselves make a life better, and vices, such as malice, make it worse. This claim need not hold that virtue is the only intrinsic good. It can and in my view should recognize other perfectionist goods, such as knowledge, and even welfarist goods, such as pleasure. But it holds that among the intrinsic goods is moral virtue and among the evils is vice.

Given this perfectionist view, a natural question is how the value of virtue compares with that of other goods, such as pleasure and knowledge. Is virtue, as some claim, the greatest intrinsic good, even infinitely more valuable than pleasure or knowledge? Is vice in a similar way the greatest evil? Or are virtue and vice each just one intrinsic value among others, sometimes outweighing other values and sometimes outweighed by them? I shall examine this comparative question. Against what I think is the most widespread view, I shall argue that, far from being the greatest good, virtue is a lesser good in the following sense: the value of a virtuous response to a good or evil object is always less than the value, either positive or negative, of that object. In a similar way, vice is a lesser evil. It is not that every in-

---

*. I presented an earlier version of this paper at the University of Maryland and of its last part at the University of Calgary and the Canadian Philosophical Association Annual Congress. For helpful comments I thank the audiences on those occasions, as well as Samuel Kerstein, Dennis McKerlie, Jamie Mayerfeld, Derek Parfit, Donald Regan, Michael Slote, Christine Swanton, Terry Teskey, and the students in my seminars at the University of Calgary. I am also grateful for support from the Social Sciences and Humanities Research Council of Canada.

0022-362X/98/9504/181–203

stance of virtue or vice has less value than any other good or evil; on the contrary, it may outweigh many goods and evils. But it always has less value than the specific good or evil that is its object. Although important additions to the list of goods and evils, virtue and vice have in this sense a subordinate comparative status.

The question of virtue's comparative value can arise given any account of what virtue is, but it is most usefully discussed with a specific account in view. Let me therefore sketch an account which I find attractive and which also raises the comparative question especially sharply. This account treats virtue and vice as higher-level values in a specific way. I call it the *recursive account*, because it combines two elements: a recursive characterization of intrinsic goods and evils and a definition of virtue and vice that fits this characterization.[1]

I. THE RECURSIVE ACCOUNT

The recursive characterization starts by identifying certain states other than virtue and vice as intrinsically good and evil. It does not matter exactly what these states are, but let us assume the characterization starts with the following mixed welfarist-perfectionist base clauses about intrinsic goods and evils:

(BG) Pleasure, knowledge, and achievement are intrinsically good.
(BE) Pain, false belief, and failure are intrinsically evil.

These clauses, and in fact all of the recursive characterization, are agent neutral in form. Thus, (BG) says that person *A*'s pleasure is good not just from *A*'s point of view but from that of all moral agents, so all moral agents have the same reason to pursue this good of *A*. (BE) makes a similar claim about the evil of *A*'s pain.

The characterization then supplements these base clauses with four recursion clauses assigning intrinsic values to certain attitudes to intrinsic goods and evils. The first recursion clause affirms the intrinsic goodness of loving what is good for itself:

(LG) If x is intrinsically good, loving x (desiring, pursuing, or taking pleasure in x) for itself is also intrinsically good.

---

[1] For earlier presentations of the elements of this account, see Franz Brentano, *The Origin of Our Knowledge of Right and Wrong*, Roderick M. Chisholm and Elizabeth Schneewind, trans. (New York: Routledge, 1969); G. E. Moore, *Principia Ethica* (New York: Cambridge, 1903); Hastings Rashdall, *The Theory of Good and Evil*, 2 Volumes (New York: Oxford, 1907); W. D. Ross, *The Right and the Good* (New York: Oxford, 1930); Robert Nozick, *Philosophical Explanations* (Cambridge: Harvard, 1981); Roderick M. Chisholm, *Brentano and Intrinsic Value* (New York: Cambridge, 1986); and my "Virtue as Loving the Good," *Social Philosophy and Policy*, IX (1992): 149-68.

By 'loving' *x*, (LG) means having a positive attitude or orientation toward *x*. This has three main forms. One can love *x* by desiring or wishing for it when it does not obtain, by actively pursuing it to make it obtain, or by taking pleasure in it when it does obtain. By loving *x* 'for itself', (LG) means loving *x* apart from its consequences or for its own sake, that is, loving instances of *x* because they are instances of *x*. Thus, one loves pleasure for itself if one loves instances of pleasure because they are pleasures, and one loves knowledge for its own sake if one loves instances of knowledge because they are knowledge.

To see more concretely what (LG) implies, let us combine it with the base clause (BG), which says that a pleasure felt by person *A* is intrinsically good. (LG) adds that, if another person *B* loves *A*'s pleasure for itself—for example, if *B* desires or pursues *A*'s pleasure for itself—his doing so is also intrinsically good. Since (LG)'s application iterates or recurs, (LG) also says that, if a third person *C* loves *B*'s love of *A*'s pleasure for itself—for example, if *C* is pleased by *B*'s desire for *A*'s pleasure—this, too, is intrinsically good. To take another example, (BG) says that knowledge is intrinsically good. (LG) adds that pursuing knowledge for its own sake is intrinsically good, as is being pleased to be pursuing knowledge.

Parallel to (LG) is a second recursion clause affirming the intrinsic evil of loving for itself what is evil:

> (LE) If *x* is intrinsically evil, loving *x* for itself is also intrinsically evil.

Combined with (BE), (LE) says that person *B*'s desiring or pursuing *A*'s pain for itself is intrinsically evil, as is *B*'s taking pleasure in *A*'s pain. If a third person *C* loves *B*'s love of *A*'s pain for itself—for example, by taking pleasure in *B*'s malicious pleasure—that, too, is intrinsically evil.

Two final recursion clauses concern the contrary attitudes of hating goods and evils:

> (HG) If *x* is intrinsically good, hating *x* (desiring or pursuing *x*'s not obtaining, or being pained by *x*'s obtaining) for itself is intrinsically evil.
>
> (HE) If *x* is intrinsically evil, hating *x* for itself is intrinsically good.

Combined with (BG), (HG) makes it intrinsically evil for *B* to desire or seek for itself the destruction of *A*'s pleasure or to be pained by *A*'s pleasure. Combined with (BE), (HE) makes it intrinsically good for *B* to be sympathetically pained by *A*'s pain—to feel compassion for *A*'s pain—or to desire or try to relieve it.

Behind the recursion clauses (LG)-(HE) is a comparatively simple idea: that appropriate attitudes to goods and evils are intrinsically good and inappropriate attitudes intrinsically evil. The appropriate attitude to an intrinsic good is to love it for its own sake, and to an intrinsic evil is to hate it. The recursion clauses therefore make these attitudes intrinsically good and their contraries intrinsically evil.

On its own, the recursive characterization says nothing about moral virtue and vice. But this omission is easily remedied by adding a definition of the virtues and vices which fits the characterization:

> (DV) The moral virtues are those attitudes to goods and evils which are intrinsically good, and the moral vices are those attitudes to goods and evils which are intrinsically evil.

Combined with (LG)-(HE), (DV) equates the virtues with loving intrinsic goods and hating intrinsic evils, and the vices with loving evils and hating goods. More specifically, (DV) equates the virtues and vices with all the higher-level goods and evils in a multilevel theory of value. The recursive characterization starts by identifying certain base-level goods and evils, namely, those of (BG) and (BE), and then adds an infinite series of higher-level goods and evils, each consisting in an appropriate or inappropriate attitude to a lower-level good or evil. (DV) equates the moral virtues with all the higher-level goods in this multilevel theory and moral vice with all the higher-level evils. The virtues and vices are intrinsic goods and evils that are made such by their orientation to other goods and evils.[2]

I find this recursive account of virtue and vice attractive. As my brief presentation of it has shown, it counts a benevolent desire for another's pleasure and compassion for his pain as virtues and malice toward him as a vice. When the account is elaborated and refined, it illuminates a wide range of further virtues and vices. The account fits naturally in a consequentialist framework, since it analyzes virtue and vice in terms of the central consequentialist properties of intrinsic goodness and evil. For those who accept consequentialism, therefore, it can give a complete account of virtue and vice. Even those who do not accept consequentialism can see it as giving a partial account of virtue and vice, one appropriate to the part of morality concerned with intrinsic goods and evils. They may want to supplement the account with further claims about nonconsequentialist forms of virtue, but can accept it as analyzing an important range of virtues and vices.

---

[2] Unlike some definitions, (DV) finds virtue and vice in occurrent attitudes rather than in longer-lasting dispositions or traits of character. There are three reasons for preferring this approach. First, any disposition relevant to virtue can only be identified as a disposition to have certain occurrent attitudes—for example, as a disposition to have the desires, perform the actions, and feel in the ways valued by some part of the recursive characterization. Second, when it is considered apart from the occurrent attitudes in which it can issue, any such disposition has minimal value. As Aristotle says, the mere possession of a virtuous disposition, as in someone asleep or otherwise inactive, counts for little. Finally, though one can hold that occurrent attitudes have more value, or even have value only, when they issue from a virtuous disposition, this is in my view a mistake. Imagine that one person feels compassion that issues from a long-lasting disposition while another feels an otherwise identical compassion but only once, without its reflecting anything permanent in his character. If we set aside any (minimal) value the first person's disposition has in itself, and also set aside the values in the other feelings and actions to which this disposition will lead at other times, is the first person's compassion better than the second's? I see no reason to believe this, and therefore no reason to find dispositions specially important for virtue and vice as intrinsic values.

The account also raises immediately the question about the comparative values of virtue and vice. The account implies that virtue and vice cannot be the only intrinsic values; their goodness requires, by definition, other values to which they can be intentionally related. So how do the values of virtue and vice compare with those of base-level goods and evils, such as pleasure, pain, and knowledge? How great a good is virtue, and how great an evil is vice?

II. VIRTUE AS A LESSER GOOD

Some philosophers hold, high-mindedly, that virtue is the greatest intrinsic good. The strongest such view holds that virtue is infinitely greater than other goods, or has lexicographic priority over them. W. D. Ross affirms part of this view when he writes that "*no* amount of pleasure is equal to any amount of virtue...; in other words, that while pleasure is comparable in value with virtue (i.e. can be said to be less valuable than virtue) it is not commensurable with it, as a finite duration is not commensurable with infinite duration" (*op. cit.*, p. 150).[3] John Henry Cardinal Newman makes a similar claim about vice, saying it would be less evil for all humankind to die "in extremest agony" than that "one soul...should commit one venial sin."[4] According to the fully general version of this lexicographic view, even the most trivial instance of virtue or vice outweighs the greatest imaginable quantity of any base-level value, such as pleasure, pain, or knowledge.

---

[3] See also Ross, *The Foundations of Ethics* (New York: Oxford, 1939), p. 275. He makes a similar claim about the infinite superiority of virtue over knowledge, his other main good, in *The Right and the Good*, p. 152.

[4] *Certain Difficulties Felt by Anglicans in Catholic Teaching*, Volume 1 (London: Longmans, 1901), p. 240; quoted in Derek Parfit, *Reasons and Persons* (New York: Oxford, 1984), p. 49.

But the lexicographic view has extremely implausible implications. Consider two possible worlds.[5] In the first world, natural conditions are benign and people therefore lead very pleasant lives, but they are all entirely self-concerned, caring about no one's enjoyment but their own. In the second world, conditions are less favorable, which gives people less pleasure, but they are benevolent and sympathetic toward each other. If we believe that virtue is intrinsically good, we shall say the second world can be better than the first, as seems right to me. But the lexicographic view implies that the second world is better even if it contains vastly less pleasure—its inhabitants suffer agony while those in the first world are blissful—and only slightly more virtue—one person feels a tiny bit more compassion. Ross and Newman are committed to accepting this implication, but I do not see how they can. Or consider a single life, that of a virtuous person who suffers agonizing pain. The lexicographic view implies that this life is overwhelmingly good with only an insignificant element of evil. But surely most of us think this life is on balance undesirable, one it would be better not to live.

The lexicographic view has equally implausible implications for cases involving action. As Hastings Rashdall asks in a penetrating critique, must we always direct our charitable contributions to people's moral improvement rather than to their education or material comfort, or at least do so when the two conflict (*op. cit.*)? If the only way to save twenty innocent people from intensely painful torture is by bribing a corrupt official and thereby further damaging his character, is bribing the official wrong? Or consider a case involving just the agent's virtue. A person contemplating a career such as nursing or surgery may know that in it she will do much to relieve others' pain. But she may also rightly fear that the constant exposure to pain will harden her character, changing her from someone who feels deeply for others' pain to one who handles it efficiently but without accompanying emotion. If virtue had infinite comparative value, it would be wrong for this person to take up nursing or surgery whatever the benefits to other people. But surely that is absurd (*op. cit.*, Volume 2, pp. 41-47). It may be objected that an unemotional nurse can still act virtuously in her career, since she can still choose her work for its own sake or for the good it does to others. But if compassionate feeling is an additional form of virtue, as it surely is, and its loss would make her life on balance less virtuous, the lexicographic view still says her choosing the career would be wrong. We can imagine a more radical example in which the exposure to suffering hardens her character further, so she no longer chooses her work for others' sake but does it only as a job or for the money it provides. (Those familiar with careers such as nursing and surgery will know this example is anything but unrealistic.) In this case, and in fact if there is even the slightest probability of this more radical outcome, the lexicographic view says this person should not take up nursing whatever the benefits to others.

---

[5] Compare *The Right and the Good*, pp. 134-35.

Despite rejecting the lexicographic view, Rashdall still holds that virtue is the greatest good: "It seems to me perfectly clear that the moral consciousness does pronounce some goods to be higher, or intrinsically better than others; and that at the head of these goods comes virtue" (*op. cit.*, Volume 2, p. 37). But it is not clear that this nonlexicographic view is possible. Consider two nonmoral goods, such as pleasure and knowledge. If neither is lexicographically prior to the other, some very intense pleasures outweigh some trivial items of knowledge and some very important knowledge outweighs some minor pleasures. But if some instances of each good outweigh some of the other, how can we say that either good is *in general* greater? If we have measured both goods on cardinal scales, we may have a formula for converting between these scales, and this formula may make, say, two units of pleasure equal to one of knowledge. But this 2:1 ratio depends entirely on our choice of units on the two scales, which is arbitrary. Given different units there could be a 2:1 ratio in the other direction. So if neither of two goods is lexicographically prior to the other, how can we make any comparison in the abstract between them?

This difficulty does not arise for the specific good of virtue, at least as understood by the recursive account. This is because every instance of virtue is, according to that account, specially connected to another good or evil that is its object. Every lowest-level virtue, to start, has as its object an instance of a base-level value, such as pleasure, pain, or knowledge. It is a desire for a pleasure, a pain at a pain, or something similar. We can therefore ask how the value of the virtue compares to that of its specific base-level object. If the virtue's value is always greater than its object's, there is a clear sense in which this form of virtue is a greater good than pleasure or knowledge; if its value is always smaller, the virtue is a lesser good. The same approach can be extended to higher-level virtues. Every such virtue is likewise connected to a specific base-level value, through the lower-level attitude that is its object. We can therefore assess the value of virtue as a whole by asking how in general the value of a virtuous attitude compares to that of its object. If the attitude's value is always finitely greater, there is a restricted but clear sense in which virtue is a nonlexicographically greater good. It may be that for every instance of virtue there are some instances of pleasure and knowledge that are better than it and some that are less good. But if there is a constant relationship between the value of the virtue and that of the specific base-level state to which it is connected, this relationship can underwrite a general claim about the comparative value of virtue. A similar relationship can underwrite claims about the comparative evil of vice. I think a charitable reading of Rashdall's nonlexicographic view takes it to rest on this kind of restricted comparison. He thinks virtue is a greater good and vice a greater evil because he thinks any virtuous or vicious attitude, though outweighed by some base-level values, has more goodness or evil than the specific base-level state that is its direct or indirect object.

Some philosophers argue that virtue is a greater good on the ground that virtue alone is morally good, whereas pleasure, knowledge, and the like are nonmorally good.[6] (This is also sometimes given as a reason for the lexicographic view.) But this argument is not persuasive. If it assumes that moral goodness is a distinct property from ordinary intrinsic goodness, it cannot hold that virtue has only that property and no other. If we can compare the values of the two worlds described above, the pleasure and virtue they contain must have the same kind of goodness. Why should the virtue's having a distinct property of moral goodness also give it more ordinary goodness? If the former property is truly distinct, why should it bear on the latter? In any case, I do not think it best to treat moral goodness as a distinct property. The simpler view is that moral goodness is just intrinsic goodness—the very same property—when had by certain objects, namely, attitudes evaluated in relation to their objects. It is just the goodness of higher-level rather than base-level goods.[7] On this view, however, the claim that virtue alone is morally good supports no substantive view about virtue's comparative worth. To establish that worth, we must compare virtue directly with other values, such as pleasure, pain, and knowledge.

---

[6] See, for example, J. L. A. Garcia, "The Primacy of the Virtuous," *Philosophia*, XX (1990): 69-91, here pp. 78-79; see also his "Goods and Evils," *Philosophy and Phenomenological Research*, XLVII (1987): 385-412, here pp. 407-09.

[7] This simpler view is that of Ross and of William K. Frankena. See *The Right and the Good*, p. 155; and Frankena, *Ethics* (Englewood Cliffs, NJ: Prentice-Hall, 1973, 2d ed.), p. 62.

In my view, this comparison supports, not Rashdall's view that virtue is a greater good, but the contrary view that virtue and vice are lesser values. This view is expressed in the following comparative principle about attitudes and their objects:

> (CP) The degree of intrinsic goodness or evil of an attitude to *x* is always less than the degree of goodness or evil of *x*.

According to (CP), the intrinsic goodness of loving a good is always less than that of the good, as is the intrinsic evil of hating it. Desiring a pleasure is less good than the pleasure is, and being pained by the pleasure is less evil. Similarly, the evil of loving an evil is always less than that of the evil, as is the goodness of hating that evil. It is not, to repeat, that every virtuous or vicious attitude has less value than any instance of another good or evil, such as pleasure or pain. A desire for the immense pleasure of others, though less good than that pleasure, can be better than a mild pleasure of one's own. But every attitude does have less value than its specific object. Virtue and vice are nonlexicographically lesser values in the sense that they are always finitely less good or evil than the specific base-level states to which they are intentionally related.

On the simplest versions of (CP), every attitude has a maximum value that is the same fraction (for example, one half) of its object's value. No matter what the attitude's intensity, its goodness or evil cannot be more than one half its object's goodness or evil. In thinking about (CP), we need not commit ourselves to any such mathematically precise formulation, which in any case is unrealistic. (I doubt that any intrinsic values can be measured with this degree of precision.) But it will help to illustrate (CP) if we imagine a version on which attitudes have at most one half their objects' value.

There are several powerful arguments for (CP). The first, which derives from G. E. Moore, considers the combination of an intrinsic evil and a virtuous response to it—for example, one person's pain and another's compassion for that pain (*op. cit.*, p. 219). If the value of a virtuous attitude were greater than that of its object, the combination of a pain and compassion for it would be on balance good, and better than if there were no pain and no compassion. But this, Moore claims, is unacceptable. The compassion is indeed good, and makes the situation better than if there were only pain and no compassion, but it cannot outweigh or justify the pain. Given a choice between creating a world containing only pain and compassion for it and creating nothing, it is surely best to create nothing. But if so, if a combination of pain and compassion for it is always on balance evil, then the positive value of virtuously hating an evil must be less than the negative value of the evil.

Or consider the combination of an evil attitude and a higher-level hatred of it—for example, a malicious desire to hurt another and shame at that desire. Since malice is evil by (LE), feeling pained by or ashamed of one's malice is good by (HE). This implication is attractive: that shame at a vicious desire is virtuous. But it would not be attractive to hold that feeling malice and shame at it is better than not feeling the malice at all. This conclusion would follow if the value of an attitude were greater than that of its object, but it does not follow given (CP).

A related argument concerns the combination of a good and vicious hatred of it—for example, person *A*'s pleasure and *B*'s envious wish that *A* were not experiencing that pleasure. *B*'s malicious envy is evil, but it surely cannot be so evil as to outweigh *A*'s pleasure or by itself make that pleasure undesirable. This conclusion is, again, avoided by (CP).

These initial arguments concern only some attitudes, namely, virtuous hatreds of evil and vicious hatreds of good. It is therefore possible to accept them while holding that other attitudes, such as virtuous loves of good, have more value than their objects. But it is surely simplest to hold that the relation between attitudes and their objects is uniform, so a conclusion like Moore's generalizes to all of (CP).

There are further arguments for (CP). One such argument concerns the value of very high-level attitudes. Given their recursive form, both (LG) and (HE) generate infinite hierarchies of intrinsically good attitudes. For example, (LG) values not only the love of pleasure but also the love of the love of pleasure, the love of the love of the love of pleasure, and so on to infinity. The clauses' recursive form is one of their attractive features. We have a better explanation of why *B*'s desire for *A*'s pleasure is good if we can say that love of any good is good, and not just the love of pleasure. It is also plausible in itself to hold that very high-level attitudes are good. But it would not be plausible to hold that these attitudes are very good or that we humans should spend much time trying to form them. (CP) avoids these implications. If the value of an attitude is always less than that of its object, say, by one half, the values of progressively higher-level attitudes get progressively smaller and diminish toward zero. Although *B*'s desire for *A*'s pleasure may have significant value, *C*'s desire for that desire of *B*'s has less value, and *D*'s desire for that desire of *C*'s even less. Very high-level attitudes still have some value, and it is a merit in, say, God that he can form them without detracting from his lower-level attitudes. But their value is sufficiently small that beings with our limited capacities should primarily work at developing lower-level virtues.

A third argument for (CP) invokes what I call the *proportionality view* about the division of virtuous love. Since we humans cannot love all good things with infinite intensity, the question naturally arises of how it is best to divide our love between different good objects. The most appealing answer is: in proportion to the objects' degrees of goodness. If x is better than y, it is best to care more intensely about x but still to some degree about y. More specifically, if x is n times as good as y, it is best to care n times as intensely about x and $1/n$ times as intensely about y.[8] In this initial use, the proportionality view takes us from claims about the comparative values of objects to conclusions about the best divisions of love. But we can equally well use it in the opposite direction, saying that, if it is best to love x more than y, then x must be better than y. There are several instances of this argument for (CP).

Imagine that a teacher works to instil knowledge in a student from a genuine desire for the student's knowledge, and that as a result the student acquires knowledge. If you learn of these facts, which should you be more pleased by, the student's knowledge or the teacher's virtuous pursuit of knowledge? Surely, you should be more pleased by the student's knowledge; it is the point of the exercise. Or imagine that you can produce only one of these goods but not both. Right now an uncaring teacher is teaching using methods that do not work, so her student is not acquiring knowledge. You can either change the teacher's attitude while leaving her methods unchanged or change her methods but not her attitude. Surely, it is best to change the teacher's methods in this case so the student acquires knowledge. But then the student's knowledge must be better. Or consider an example involving evils. If a torturer is causing a victim intense pain and also taking sadistic pleasure in that pain, surely you should be most pained by the victim's pain. If you can either interfere with the torturer's pleasure while leaving his torture machine running or secretly disconnect the machine, you should disconnect the machine. If so, the evil of the torturer's pleasure in pain must be less than that of the pain.

---

[8] For a fuller account of the proportionality view, see my "Self-Interest, Altruism, and Virtue," *Social Philosophy and Policy*, XIV (1997): 286-307. Note that divisions of love that are slightly disproportionate are not evil. They are still good, just not ideally so.

Or consider attitudes to one's own virtue. What should the teacher in the example above have as her main motive for teaching, that she will thereby act virtuously or that through her teaching her student will gain knowledge? What should she be most pleased by afterward, her own good action or its effect? Many writers hold that there is a moral failing, called variously moral self-indulgence, narcissism, or priggishness, that consists in caring too much about one's own virtue. Bernard Williams[9] says an agent is morally self-indulgent when "what the agent cares about is not so much other people, as himself caring about other people.... [A] person may act from generosity or loyalty...and not attract the charge of moral self-indulgence, but that charge will be attracted if the suspicion is that his act is motivated by a concern for his own generosity or loyalty" (*ibid.*, p. 45). Williams does not hold that *any* concern for one's own virtue is a failing; self-indulgence consists only in caring "disproportionately" about one's virtue, or having it be "more important" in one's motivation than a concern for other people (*ibid.*, pp. 47, 45).[10] He therefore allows that some concern for one's virtue, provided it is a lesser concern, is consistent with ideal attitudes. His view therefore fits perfectly with (CP). Because the teacher's virtuous action is intrinsically good, it is indeed appropriate for her to desire and be pleased by it. But if her action is less good than the knowledge it produces, what is best is for her to love her virtue less than she does the knowledge. If she cares more about her virtue, she divides her love disproportionately, which is precisely a moral failing.

Not all philosophers who condemn self-indulgence can do so on the basis of (CP). Ross,[11] most notably, recognizes the "self-absorption" of an excessive concern with one's own character, but also holds that virtue has infinite value compared to other goods. The question, however, is whether a position like Ross's is consistent. It is surely plausible that given two goods, one greater than the other, it is best to care more about the greater than about the lesser. This is the core of what I have called the proportionality view about division. If we accept this view, however, and if virtue is even a finitely greater good, it follows that it is best to care more about one's virtue than about its effects. It is a commonplace of contemporary ethics that self-indulgence is a moral failing and that truly virtuous agents are not motivated primarily by thoughts of their own virtue. What is less often noticed is the tension between these claims and the still widely held view that virtue is the greatest among goods. In fact, the best way to explain why self-indulgence is a failing is to hold, with (CP), that virtue is a lesser good.

---

[9] "Utilitarianism and Moral Self-Indulgence," in his *Moral Luck* (New York: Cambridge, 1981).

[10] See also Noah M. Lemos, "High-minded Egoism and the Problem of Priggishness," *Mind*, XCIII (1984): 542-58, here p. 550.

[11] *Aristotle* (New York: Macmillan, 1949), p. 208.

A final, more abstract argument for (CP) concerns virtue's status as a dependent intrinsic good. On the recursive account, virtue consists in certain attitudes to other goods and evils and therefore presupposes other goods and evils. It cannot be the only intrinsic value, but requires there to be base-level values to which it responds. It does not follow that virtue cannot be intrinsically good—the recursive account shows that it can—but a weaker conclusion does seem appropriate. If virtue is an attitude to other goods and evils, it seems fitting that it should be a lesser good, one whose value cannot exceed that of its object. Something secondary in the source of its value should likewise have secondary weight. There is no logical compulsion about this argument; it is perfectly consistent to hold that virtue is a response to other goods but, as that response, has infinitely more value. Nonetheless, it seems in a looser way appropriate that a good defined by reference to other values should be subordinate in its comparative standing.

I have given four types of argument for (CP): (1) about certain combinations of an object and an attitude, such as pain and compassion for it; (2) about very high-level attitudes; (3) about the best divisions of love; and (4) about virtue's status as a dependent good. Together, these arguments strongly support (CP). (CP) holds that while any instance of virtue or vice may outweigh some instances of other goods or evils, it always has less value than the specific good or evil to which it is intentionally connected. This claim is one we can consider only given something like the recursive account of virtue and vice. If we treated the virtue concepts as primitive concepts rather than as connoting attitudes to other goods or evils, we could consider only lexicographic claims about virtue's general comparative worth. If we rejected those claims, we could not say anything in the abstract about virtue's rank among values. But if we take every instance of virtue and vice to be directed at another good or evil, we can compare its value to that of that specific good or evil. In this comparison, I have argued, virtue and vice are lesser intrinsic values.

III. A LESSER GOOD: DIFFICULTIES

There are, then, powerful reasons to accept the comparative principle (CP). But there are also difficulties about (CP) that a full discussion must address. Let me consider three such difficulties.

The first difficulty concerns Moore's argument about pain and compassion. (CP) implies that a combination of a pain and one person's compassion for it is always on balance evil. But what if two, twenty, or a hundred people feel compassion? Are the resulting combinations still on balance evil? I think our initial response is that they are. Our judgment may not be as confident here as in the case involving one person's compassion, but I think many of us will say that given a choice between creating pain plus a hundred people's compassion and creating nothing, it is best to create nothing. But it is hard to see how the recursive account even with (CP) can endorse this judgment. If each of the hundred people's compassion is good, then even if each is less good than the pain is evil, can the sum of their goodness not outweigh the pain? Can enough virtuous attitudes to pain not make for on-balance positive value? This conclusion would not follow if the value of extra people's compassion diminished the more people felt compassion, but there seems no reason to believe it does. That others are sympathizing with someone's pain makes it no less appropriate for me to do so and should therefore make it no less good. We may think the value of extra compassion diminishes if we consider compassion instrumentally, as helping to console the victim. (If ninety-nine people are already sympathizing with him, he will not take much more comfort from a one-hundredth.) But we do not think this way if we consider the compassion by itself. Then the value of each person's compassion seems the same, and this implies that enough people's compassion can outweigh any pain.

This difficulty weakens the Moorean argument for (CP). That argument claims that by accepting (CP), the recursive account can capture our intuitive judgments about combinations of pain and compassion; if some such judgments are not captured, the support given (CP) is less. But it is not clear in what direction the difficulty points. It is certainly not a reason to hold that virtue is a greater good; that would only make the problem worse. If we cannot hold that the value of extra compassion diminishes, we may have to accept the conclusion about numbers, difficult though that may seem. We want to hold that virtue is intrinsically good; we also want to hold that in some cases, like that of the two worlds described above, virtue can outweigh base-level values, such as pleasure and pain. Given these assumptions, however, and adding that the value of extra people's compassion is constant, it seems unavoidable that some number of people's compassion can outweigh any pain. Although it is initially unsettling, the conclusion seems to be one we have to accept.[12]

The second difficulty concerns (CP)'s scope. As originally formulated, (CP) applies indiscriminately to all attitudes, saying they all have less value than their objects. This claim is compelling for attitudes whose objects are great goods or evils—for example, are intense pleasures or pains. But it is less so for attitudes whose objects are trivial goods or evils. Imagine that *B* goes out of his way to provide from benevolence a small pleasure for *A*, say, by doing *A* a small favor. May in this case *B*'s benevolence not have more value than the pleasure it produces? (Of gift giving we say, "It is the thought that counts.") Or imagine that *B* takes special trouble to cause *A* a small annoyance or pain. May his petty malice not be worse than its effect? In general, may attitudes to trivial objects not have greater value than their objects?

If we were persuaded by these examples, we would have to restrict (CP) so it applies only to attitudes whose objects have significant value. More generally, we could revise (CP) so the ratio between the value of an attitude and that of its object is not constant, as in the simple versions discussed above, but changes continuously with the object's value. When the value of the object is high, the value of the attitude is a small fraction of it. As the object's value declines, the fraction gets larger; and when the object is trivial, the attitude's value is greater. Instead of a constant ratio, such as one half, between the value of an attitude and that of its object, there are different ratios for objects of different values.

Unfortunately, this revised (CP) is incompatible with the proportionality view about the best divisions of love. That view, recall, holds that, if x is n times as good as y, it is best to love x n times as intensely as y. The simplest justifications of this view assume that, if x is n times as good as y, loving x with a given intensity is always n times as good as loving y with that intensity.[13] This last assumption can be maintained if the value of an attitude is always a constant fraction of its object's value, as on the simple versions of (CP). But it cannot be maintained given the revised (CP). The revised principle therefore requires abandoning the proportionality view in favor of one on which the best division of love between *x* and *y* is not strictly proportioned to their values but gives slightly more love to the lesser good *y* than its degree of goodness would suggest. Some may be willing to accept this departure from strict proportionality and may therefore adopt the revised (CP). But I find the original proportionality view highly appealing and therefore prefer to retain the simpler (CP). This requires accepting what some may find counterintuitive, that even attitudes to trivial goods and evils have less value than their objects. (In gift giving, it is the gift that counts.) So each of the available options is in some respect questionable. But whichever option we choose, the recursive account will retain the core claim of (CP): that for objects above a threshold of goodness or evil, the value of any attitude to them is less than their own.

---

[12] We may also have to accept a parallel conclusion about time, namely, that a sufficiently long-lasting compassion can outweigh any pain. This conclusion may be mitigated to some extent if we hold that the degree of compassion appropriate to a pain, and therefore the value of that compassion, diminish as the pain recedes into the past. But it is not clear that this condition will always prevent long-lasting compassion for a short-lived pain from being more good than the pain itself is evil.

[13] See the discussions of the "asymptotic" and "optimality" views in my "Self-Interest, Altruism, and Virtue."

The most important difficulty for (CP) concerns the value on balance of certain vices of loving evil, such as sadistic pleasure in another's intense pain. Imagine that a torturer takes pleasure in the pain he causes his victim. The recursive account says, through its recursion clause (LE), that his pleasure is intrinsically evil as an instance of loving evil, or of vice. But it also says, through its base clause (BG), that his pleasure is intrinsically good as an instance of pleasure, or as having the pleasantness that makes all pleasure sensations as such desirable. The torturer's pleasure is therefore intrinsically evil as pleasure in pain but intrinsically good as pleasure. So is it on balance intrinsically good or intrinsically evil? We surely want it, intuitively, always to be on balance evil, so its presence always makes a situation worse. In fact, its ability to make such pleasure intrinsically evil is a central attraction of the recursive account as against purely subjective views, such as hedonism. But whether sadistic pleasure is on balance evil depends on the comparative values of its qualities of vice and pleasantness. Only if the pleasure's evil as vice always outweighs its goodness as pleasure will it always be on balance evil. How can this be, given (CP)? Imagine that (CP) makes the maximum value of an attitude always one half that of its object, and that equal units of intensity of pleasure and pain have equal value; then, if the torturer takes two units of pleasure in his victim's ten units of pain, his pleasure may be on balance evil. It can have more than two units of disvalue as vice and therefore make the overall situation worse. But if he takes six units of pleasure in his victim's ten units of pain, his pleasure has more goodness as pleasure than its maximum five units of vice, and makes the situation better.

If it is sufficiently intense, a sadistic pleasure can be, counterintuitively, on balance good.[14]

This difficulty would not arise given the lexicographic view, which makes virtue and vice infinitely more valuable than pleasure or pain. Ross cites this fact in arguing for the view: "If the goodness of pleasure were commensurable with the goodness or badness of moral disposition, it would be possible that such a pleasure [that is, a pleasure of cruelty] if sufficiently intense should be good on the whole. But in fact its intensity is a measure of its badness."[15] As we have seen, however, there are compelling reasons to reject the lexicographic view. Once virtue and vice have just finite value compared to their objects, let alone lesser value, it seems that some instances of pleasure in intense pain can be on balance good.

Can the recursive account avoid the implication that sadistic pleasure can be on balance good? One way is to deny that pleasure as such is intrinsically good, or belongs among the base-level values identified by (BG). This antihedonist view can allow that pleasures with good objects are good; they are still virtues as valued by (LG). But it denies that pleasure considered independently of any object, or in itself, is good.

This antihedonist view certainly avoids the claim that pleasure in evil can be on balance good; if doing so were a necessity, we might have to adopt it. But the view has many other counterintuitive implications. For example, it denies that nonintentional pleasures, such as those of suntanning or eating ice cream, are good. Although not completely unimaginable, this is an extreme step. Surely, most of us believe that simple bodily pleasures have some intrinsic worth. A less sweeping antihedonism denies only that intentional pleasures are good as pleasures: unstructured bodily pleasures have value, and only pleasures that something is the case do not. But this view still denies, implausibly, that pleasures with neutral intentional objects are good. Consider a sports fan's pleasure, which may be very intense, that his local team has won a championship. If his feeling involves the same quality of pleasantness as the pleasure of suntanning, should it not likewise be good? Nor do only pleasures with neutral objects seem good as pleasures; those with good objects do as well.

---

[14] This difficulty is nicely illustrated in Frankena's *Ethics*. He claims both that malicious pleasure is always on balance evil and that every pleasure has some value as pleasure (pp. 89-90). But he does not explain why, as this combination of claims assumes, the moral evil of malicious pleasure must always outweigh its hedonic goodness as pleasure.

[15] *The Right and the Good*, p. 151.

Imagine that *A* lives in a world with many very good objects in which he takes appropriate pleasures, whereas *B* lives in an evil world by which he is appropriately pained. Do we not think that *A*'s state, though no more virtuous than *B*'s, is on balance better because his virtue can take the form of pleasure?

It may seem that the solution to these difficulties is obvious: deny only that pleasures with evil intentional objects are good as pleasures. Nonintentional pleasures are good as pleasures, as are pleasures with neutral or good objects. Only pleasures in evil objects, such as sadistic pleasures, lack value as pleasures. Their orientation to evil has two effects, making these pleasures evil as attitudes and cancelling their goodness as pleasures. But attractive though it may seem, this moralized antihedonism would require major changes in the structure of the recursive account and even then cannot be captured in its entirety.

In the original recursive account, the base clauses (BG) and (BE) are prior to and independent of the recursion clauses (LG)-(HE), identifying their nonmoral goods and evils without reference to any subsequent claims about good and evil attitudes. The base clauses are also independent of each other, with neither restricted by the other's claims. But these structural features exclude the moralized antihedonism just proposed. To say that sadistic pleasure has no value as pleasure is to make a claim about base-level values. To say it has no value because it is directed at evil is to ground that claim in something recursive; and in the original recursive account, no base-level claim can depend on any recursive claim.

Because of this, capturing the proposed antihedonism would require changing the account's structure, first by abandoning the independence of (BG) from (BE). A revised account starts with the original (BE), which says that all pain, false belief, and failure are evil. It then follows (BE) with a version of (BG) that is restricted by (BE), one saying that pleasure is good *except when* it is directed at the evils identified by (BE), that is, except when it is directed at pain, false belief, or failure. This revised (BG) implies, as desired, that sadistic pleasure has no value as pleasure; when the recursion clause (LE) is added, sadistic pleasure will always be on balance evil. The revised (BG) also allows that pleasures with no, neutral, or good objects are good. So it has at least initially the right kind of implications. But it does not capture all the proposed antihedonism, since it does not imply that pleasures with higher-level evil objects lack value as pleasures. Imagine that while *B* takes sadistic pleasure in *A*'s pain, *C* takes a higher-level pleasure in *B*'s sadistic pleasure, enjoying the fact of *B*'s sadism. Surely, if *B*'s pleasure lacks value as pleasure, *C*'s should as well; it, too, is directed at evil. But a (BG) restricted only by (BE) cannot endorse this claim, since it denies value only to pleasures in base-level evils. If it is to yield the claim, the recursive account must be revised further to abandon the priority of all base-level over all recursive claims. More specifically, it must interpose the recursion clause (LE) about hating evils between (BE) and (BG). Then (BG) can be doubly restricted, saying that pleasure is good except when it is directed at the evils identified by either (BE) or (LE). This second revision makes *C*'s pleasure in *B*'s sadistic pleasure lack value as pleasure, as desired. But it still does not capture all the proposed antihedonism. Imagine that *B* hates a good of *A*'s—for example, is enviously pained by *A*'s pleasure—and that *C* takes a higher-level pleasure in *B*'s envious pain, delighting in it because it is envious. Here, *C* loves an evil attitude of *B*'s, as in the previous example, but this attitude is now one of hating good rather than loving evil. Since the envious pain that *C* loves is evil by recursion clause (HG), *C*'s pleasure in it should not on our current proposal have value as pleasure. But this last claim cannot be captured by any changes in the recursive account. The clause that makes *B*'s envious pain evil, (HG), concerns the hatred of goods, and unlike (LE), which presupposes only base-level evils, (HG) requires a prior identification of base-level goods. But this characterization is supposed to be given by the very base clause (BG) that we are now trying to restrict! Capturing the desired claim about pleasure in envious pain therefore involves a circularity: (BG) is to be restricted by reference to a recursion clause (HG) that in turn requires a completed (BG) before its own claims have substance. Changing the structure of the recursive account can capture some of the view that pleasures with evil objects lack value as pleasures, but it cannot without circularity capture all of it.

The original structure of the recursive account, with independent base clauses prior to any recursion clauses, is attractive in its simplicity. Because of this, any change that complicates the account's structure is in that respect unattractive. It is especially so if the change does not result in a fully motivated or principled view. Surely, what is plausible here is only the fully general view that all pleasures with evil objects lack value as pleasures; any more partial view is unacceptably ad hoc. So, if the recursive account cannot capture all the proposed moralized antihedonism, because it cannot say that pleasures whose objects are made evil by (HG) lack value as pleasures, that antihedonism should be abandoned.

Where does this leave us? In my view, there is only one principled way for the recursive account to hold that all pleasure in pain is on balance evil, and that is to deny that pleasure as such is a base-level good. This is not an unimaginable move, but it has the extreme implication that innocent pleasures, such as those of suntanning or supporting a sports team, have no intrinsic value. If we cannot accept this implication, we must grant that even pleasures in pain have value as pleasures and can therefore sometimes be on balance good. But this may not in the end be an unacceptable claim.

First, the claim parallels one that is positively attractive. The recursive account holds through base clause (BE) that compassionate pain, that is, pain at another's pain, is always evil as pain. Given (CP), it therefore also holds that compassionate pain can sometimes be on balance evil. But this is not an unattractive implication. On the contrary, it fits our practice of sometimes not revealing our hurts to our friends, to spare them the pain of sympathizing with us. We may recognize that our friends' compassion would be good as virtuous, but we seem sometimes to view its evil as pain as more significant. But if in this case the moral goodness of an attitude can be outweighed by its hedonic evil as pain, why cannot in a parallel way the moral evil of pleasure in pain be sometimes outweighed by its hedonic goodness?

Second, there are cases where it seems right to hold that pleasure in pain is on balance good. Imagine that *B* takes pleasure in some minor mishap of *A*'s, such as *A*'s slipping on a banana peel, or laughs at a slightly pointed joke at *A*'s expense. (A great many jokes, including most political jokes, turn in a slightly malicious way on some person's failings.) On the recursive account, there is something morally objectionable about *B*'s pleasure; it would be better if he got the same enjoyment from a neutral or, better, a good object. But what if he cannot do so? What if the only options given *B*'s character are his getting pleasure from slight evils for *A* and his getting no pleasure at all? Is it then best if *B* gets no pleasure? This seems an excessively prudish view.[14] It is, to repeat, unfortunate that *B* can get pleasure only in this way; it would be better if he had nonmalicious sources of enjoyment. But to say that his pleasure is on balance evil is to condemn, all things considered, what is for many people one of their readiest sources of enjoyment. The issue here is not the value of pleasures in great evils, such as a torturer's pleasure in his victim's intense pain. It is only that of pleasures in small evils, ones the love of which (CP) makes an even smaller evil. It does not seem wrong to hold that these pleasures' small evil as vices is outweighed by their greater goodness as pleasures. On the contrary, it seems right, on all but a prudish view.

---

[14] For a discussion of jokes that seems to assume this prudish view, see Ronald de Sousa, *The Rationality of Emotion* (Cambridge: MIT, 1990), chapter 11.

Pleasures in small evils, then, can be on balance good. But what about pleasures in great evils, such as the torturer's sadistic pleasure in his victim's intense pain? Is it not still counterintuitive to hold that they can be on balance good? It is, but the recursive account can agree that these pleasure are always on balance evil if it revises its treatment of pleasure as a base-level good, and in particular denies that equal units of intensity of pleasure and pain always have equal value.

The base clauses (BG) and (BE) imply that pleasures and pains are, respectively, better and worse when they are more intense, but they do not say how much they are better and worse. The simplest view says the relationships here are linear: the value of an additional unit of intensity of pleasure or pain is always the same, so pleasures and pains that are twice as intense are always twice as good or evil. This linear view is attractive for the intensity of pains, and I shall assume at least initially that the account applies it to pains. But the view is considerably less attractive for pleasures. Imagine that *A* has a choice between the certainty of a pleasure of ten units of intensity and a gamble with a .51 probability of giving him a pleasure of twenty units of intensity and a .49 probability of giving him no pleasure. The linear view says he should prefer the gamble, but I think most of us would say he should not. Or imagine that *C* can give either *A* a pleasure of twenty-one units of intensity or each of *A* and *B* a pleasure of ten units. The linear view says *C* should give *A* the twenty-one units but, again, many would disagree. To capture the judgments these examples suggest, the recursive account can abandon a linear treatment of pleasure and hold instead that the value of an extra unit of intensity in a pleasure gets smaller as the pleasure's intensity increases, diminishing asymptotically toward zero. On this asymptotic view, the second ten units of intensity in a pleasure have less value than the first, so *A* should not take his gamble and *C* should give the equal pleasures to *A* and *B*. Assuming these are attractive implications, we have independent reasons to adopt the asymptotic view; and doing so can avoid the implication that pleasures in great evils can be on balance good.

If the recursive account retains a linear view of pain, and holds that the maximum value of an attitude is always the same fraction of its object's value, it holds that the moral evil of loving a pain that is twice as intense is always twice as great. On any view, this implies that the intensity of pleasure required to outweigh this moral evil is also greater. But given the asymptotic view of pleasure, the value of increases in the intensity of a pleasure gets progressively less. This implies, initially, that many fewer cases of pleasure in pain are on balance good than given a linear view of pleasure. More importantly, it implies that some pleasures in pain never can be on balance good. Since the asymptotic view makes the value of increases in the intensity of pleasure diminish toward zero, it places an upper limit on the value of any pleasure, whatever its intensity. This implies that, if the moral evil of a pleasure in pain is above that limit, its moral evil cannot be outweighed by its hedonic goodness. Imagine that the upper limit on the value of a pleasure is ten units, and that the value of an attitude of intensity *i* is always one half that of its object. Then, if *B* takes pleasure of intensity *i* in *A*'s thirty units of pain, the value on balance of his pleasure must be negative. Its moral evil as an attitude to its object is fifteen units, which is greater than the maximum ten units of goodness it can have as a pleasure. So his vicious pleasure makes the situation worse. By combining a linear treatment of pain with an asymptotic treatment of pleasure, the recursive account can allow that some pleasures in mild pains are on balance good. This, I have argued, is a plausible implication. But it can also hold that pleasures in more intense pains, ones above a certain threshold, are always on balance evil. By treating pleasure and pain asymmetrically, it can consistently hold that pleasure in all forms is a base-level good, that attitudes always have less value than their objects, and that pleasures in great pains are never on balance good.

The more specific implications of this approach depend on exactly how fast the value of extra units of pleasure diminishes, or exactly what the upper limit on a pleasure's value is. If this limit is low, comparatively few cases of pleasure in pain are on balance good, but intense innocent pleasures have only limited value, compared to other goods. If the limit is higher, innocent pleasures have more value, but more pleasures in pain are good. It may be that no precise version of the approach is attractive in every respect: either it gives what seems too little value to innocent pleasures, such as those of suntanning, or it allows too many vicious pleasures to be good. This dilemma may be relieved to some degree if the recursive account extends its asymmetrical treatment of pleasure and pain by adopting a nonlinear view of pain contrary to the one adopted for pleasure. On this view, the value of an additional unit of intensity in a pain is not constant but gets greater as the pain's intensity increases, so pains that are twice as intense are more than twice as evil.[17] This second nonlinear view makes intense pain more evil, which given (CP) makes sadistic pleasure in it also more evil as vice; and this allows the upper limit on the value of a pleasure as pleasure to be higher while pleasures in intense pain are still always on balance evil. This doubly nonlinear approach may not entirely resolve the dilemma about good versus evil pleasures. Nonetheless, either it or the simpler asymmetrical view described first seems to offer the best available response to the difficulty about the value on balance of sadistic pleasure, allowing the recursive account to retain its original structure, affirm (CP), and still deny that the worst pleasures in evil can be on balance good. In fact, this is an additional reason to embrace the asymmetry: it offers the only way of reconciling, within a recursive framework, the overwhelmingly plausible (CP) and the intuitive view that sadistic pleasures are always on balance evil.

IV CONCLUSION

Let me summarize the main points of this paper. I have considered the comparative intrinsic values of virtue and vice, understood as perfectionist values, and have done so within a recursive account of virtue and vice that allows certain nonlexicographic claims about them to be formulated. I have argued that virtue and vice are lesser intrinsic values in the sense that their intrinsic goodness or evil is always less than that of their specific intentional object. Although instances of virtue and vice may outweigh some instances of other goods and evils, they always have less value than the specific one to which they are intentionally connected. Finally, I have considered certain difficulties for this comparative view, arguing that they do not provide good reasons to abandon it. That virtue and vice are lesser values is not without contestable implications, but it is the best available view on this topic.

THOMAS HURKA

University of Calgary

---

[17] This view is defended in Jamie Mayerfeld, "The Moral Asymmetry of Happiness and Suffering," *Southern Journal of Philosophy*, xxxiv (1996): 317-38.

BOOK REVIEWS

A Novel Defense of Scientific Realism. JARRETT LEPLIN. New York: Oxford University Press, 1997. xii + 204 p. Cloth $39.95.

The beauty of this book, as I behold it, lies in the slender basis from which it deals with a host of deep and interesting problems. Its structure is simple: it begins by showing how and why to restore truth—partial truth—to its rightful place in explaining scientific realism and progress. It deals next with theories about novelty, eliciting from its criticisms of these theories several requirements which are analyzed, substantiated, and illustrated. Then it deals with counterexamples and counterarguments against its version of realism (more precisely, minimal epistemic realism). In the process, it deftly handles a large array of issues: context of discovery and justification, instrumentalism, empirical equivalence, underdetermination, abduction, historical skepticism, auxiliaries, Goodmanesque properties, Bayesian confirmation, and, of course, the natural ontological attitude (NOA). That said, I turn to some concerns, two in particular.

I. PARTIAL TRUTH

Jarrett Leplin is admirably forthright. "I am vague by default as to how much novel success merits what level of confidence in representational success" (128). In short, he has little or no epistemic measure for partial truth. But I am less concerned about the vagueness, or absence, of the epistemic criteria by which to measure the partial truth of a theory than about the concept of partial truth itself. It is plausible to assume, says Leplin, that

$$pr(UF) > pr(PT)$$

where UF is utter falsity and PT partial truth. The reason for this inequality: "the prior probability of PT, absent S [success], is presumably lower than that of UF, if only because ways of going [utterly?] wrong would seem more profuse than ways of being [absolutely?] right" (129).

Let us begin with some distinctions. At one end, say, there is absolute truth; at the other end, there is absolute falsity. (Leplin holds both notions suspect (130).) Everything in between is partially true or partially false. But Leplin introduces the notion of utter truth and, correspondingly, one might speak of utter falsity. Let us say that all theories very close to absolute truth are utterly true, and those very close to absolute falsity are utterly false. (How would Leplin measure this very closeness?) There would still be a large left-over class of the really interesting theories in the middle that are partially true. The difficulty with Leplin's stance becomes stark. Let us say that the inequality holds, and that it also holds when AT (absolute truth) is substituted for PT; but the inequality manifestly does not hold in nearly all the other relevant, frequent, and interesting cases to which Leplin himself adverts throughout the book, namely, those cases in which one has to measure the relative distance from truth of two partially true theories. In short, under what conditions does the following inequality hold (absent S [success] from both theories)?

$$pr(PT(T_2)) > pr(PT(T_1))$$

The reasoning that the prior probability of one theory, T_1, is presumably lower than that of the other theory, T_2, if only because ways of being partially true as T_2 dictates it would seem more profuse than ways of being partially true as T_1 dictates it, evidently does not apply here. We have *no* such measure.

Consequently, it is not only disappointing, but surprising, when Leplin treats the problem as if it were a question of pragmatics (132-34). He says: "Seldom does acceptance amount to much of an epistemic commitment. It is primarily a pragmatic commitment to act on the theory *as if it were true*" (132; my emphasis). (Might Arthur Fine smile?) Why is it a question of pragmatics? We have no way of judging the absolute worth of a current theory, although we can judge the worth of a past theory in the light of the currently accepted one. How do we judge that? "Current science identifies certain interests" (134). If past theories served those interests, then we shall label them as partially true; otherwise, utterly false. The late Paul Feyerabend used to think that with the demise of a major theory, or culture, we leave unsolved a host of problems and issues, and those which were solved are forgotten or undervalued; furthermore, he thought, we do not know that the outstanding problems are not worth solving. Consequently, it might also be epistemically relevant and interesting to judge the present on the basis of the past. In the absence of such a judgment, our current assessment of both the present theory and the ones of the past are simply pragmatic. It is a pragmatic assessment in the sense of not being a partial-truth indicator. (Might Bas van Fraassen nod in approval?)

Leplin needs the notion of '—is partially truer than—', if much of his argument in the book is to succeed; for he wants to demonstrate not just that the minimal epistemic realism is true, but also that there is scientific progress. In other words, later theories in the history of science that satisfy his stringent requirements and conditions are not just simply partially true, but rather are partially *truer than* their predecessors (133). Their being partially truer than their predecessors is then supposed to explain their greater empirical success; this, in turn, is offered as an argument for minimal epistemic realism. Consequently, Leplin urgently needs an account—even a sketch—that will show how one false theory is partially truer than another false theory. Karl Popper's methodology, for over twenty-three years, has stayed wrecked on the notion of verisimilitude. Would Leplin's methodology, employing as it does partial truth, fare any better?

II. NOVELTY

Not any notion of novelty would do, if a theory is to satisfy the Leplian strictures. Thus, "two conditions for the prediction of an observational result *O* to be novel for a theory *T*" are as follows:

> *Independence Condition*: There is a minimally adequate reconstruction of the reasoning leading to *T* that does not cite any qualitative generalization of *O*.
> *Uniqueness Condition*: There is some qualitative generalization of *O* that *T* explains and predicts, and of which, at the time that *T* first does so, no alternative theory provides a viable reason to expect instances (77).

I shall not say anything about the independence condition here. The centerpiece uniqueness condition, I think, is too strong. Consider: at t_1, T_1 and T_2 are two theories in the same domain, say, physics; there is a well-established empirical result, *e*, which both theories are attempting to explain. T_1 is able to do so before T_2; at t_2,—a week, a month, a half-year later—T_2 is also able to explain it. Since its successful explanation of *e*, T_1 has not been further tested, nor is it known how new predictions are to be derived from it. At t_2, both theories are able to explain the phenomena then known. Yet the uniqueness condition allows T_2 no kudos; indeed, so strict are its standards that Leplin says that T_2 should be thought to be correct only "by chance" (64), fluke, happenstance. So: "More generally, truth is not to be attributed to a theory in explanation of its explanatory success if the result explained can also be explained another way" (64-65).

This is quite implausible. First, had we reversed the roles of the two theories, T_1 would have been judged to be correct by fluke, and T_2 would have been judged successful, partially true. Other than the time factor, there seems no warrant, in such cases, for Leplin's claim. Indeed, given Leplin's later, and large, concessions to antirealism, based on the explanationist tendencies of practicing scientists in quantum physics (182-86), this is surely unacceptable. There is an evident difference between a scientist cooking up a theory, algorithmically (158-61) and such, which Leplin justly condemns, and a scientist devising a genuine conjecture, albeit belatedly. Second, if T_2 were tested and a new consequence derived from its theoretical mechanism were experimentally confirmed, one would not be obviously wrong in saying that T_2 was not merely right by chance, it was partially truer than T_1—especially, if the latter had made a prediction, based on its own theoretical mechanism, which had then been experimentally falsified. Third, Leplin argues that should such cases arise, scientists would reserve judgment:

> But where rival theories successfully predict the same results and are not otherwise impeachable—conditions that violate the uniqueness condition—scientists are unlikely to accept any of them.... And if a theory was never believed in the first place, it is hardly incumbent on me to provide for its retention (152).

This is a bit like saying that the "Black Knight" and the "Queen's Sap" were in dead heat for the first place so we shall declare no winner. Leplin mistakes the import of the argument. The argument is not that he, Leplin, advocates the acceptance of a theory that is not acceptable, but rather that he fails to advocate acceptance of a theory that should be acceptable; or, more carefully, he advocates acceptance of one theory when it might be best to accept both. The uniqueness condition is so strong that in cases of coeval theories, it cannot help declaring a winner—given that, inevitably, one of those theories must have been proposed first.

"In this situation," says Leplin, "if we are not to credit the theory with some measure of truth, then we have *no* way to understand how the theory is able to meet this challenge. The result itself need not be considered mysterious, in that we have a theory to explain it. But this explanatory achievement is itself a mystery. Nothing but the theory's truth explains it" (100). But that is precisely what is being challenged. Look at it this way. The uniqueness condition is correct, says Leplin, because it will allow only one hypothesis as a winner; if only one hypothesis is a winner, explanation of a well-known result ceases to be mysterious. It ceases to be mysterious because the winner is declared to be true, and truth is explanatory. But mystery is precisely what is *re-introduced*, in a very plausible way, in the foregoing example. Without anything gimmicky, T_2 explains e just as well as T_1 does. Thus, if the uniqueness condition is not to be defended on the basis of an obviously circular argument, one will need a separate argument in defense of the uniqueness condition which will show why T_2 fails to be explanatory of e; in short, why T_2 is correct only "by chance." The answer, "T_1 was there first," will simply not do.

Let me demonstrate this differently since the matter is so enormously important for the book. Leplin says:

> ...the false, successful theories bear some connection or similarity to the true beyond predictive success, which explains their predictive success. Perhaps they posit similar structures or processes, and are thereby *approximately* true, or at least not *completely false*, to give the similarity a suggested if tentative description. But whatever the similarity turns out to be, it should be clear that there must be one or there simply is no explaining predictive success under circumstances of underdetermination (14-15).

This is, surely, overstated. First, Leplin is avowedly unhappy with the notion of absolute truth, as we saw; consequently, he must mean that the theory in question, T_1, against which others, for instance, T_2, are compared for similarity, is only partially true. Second, if so, he is obliged to make a much stronger claim to make his argument go through, namely, that T_1 is partially truer than T_2. We saw, too, that he has no basis to make this claim. But, third, Leplin also needs that claim to assert correctly that "the false, successful theories [such as T_2] bear some connection or similarity to the true [namely, T_1] beyond predictive success, which explains their predictive success." Consequently, fourth, the only conclusion Leplin is entitled to is that the two theories must have some structures or processes in common to explain their success and, *for all we know*, it might be T_2 that is partially truer than T_1 and the explanation of the success of T_1 may well be parasitic upon the explanation of the success of T_2.

Why is Leplin so keen on the uniqueness condition? Because if two or more theories, with vastly different theoretical mechanisms, can satisfy Leplin's independence condition and become acceptable (although, obviously, both cannot satisfy the stringent uniqueness condition), it would allow the antirealist to get a foot in the door. We would, perforce, have to say that the two or more contrary theories are equally good, or equally partially true; but not be able to say which is partially truer than which, nor be able to explain—on Leplin's realist premises, anyway—the evident success of the two theories.

If the foregoing is right, then, perhaps it has also indirectly argued for resurrecting the notion of temporal novelty, as a way of independently confirming that a hypothesis has indeed explained what it alleges to have explained: an already established fact, *e*. The old idea of welding the notion of independent testability to the notion of scientific progress might, after all, not be entirely mistaken.

HUSAIN SARKAR

Louisiana State University/Baton Rouge

NEW BOOKS

CALHOUN, LAURIE: *Philosophy Unmasked: A Skeptic's Critique.* Lawrence: Kansas UP, 1997. xii, 212 p. Cloth $25.00.

CAO, TIAN YU: *Conceptual Developments of Twentieth Century Field Theories.* New York: Cambridge, 1996. xx, 433 p. Cloth $59.95.

CAPUTO, JOHN D.: *The Prayers and Tears of Jacques Derrida: Religion without Religion.* Bloomington: Indiana UP, 1997. xxix, 379 p. Cloth $39.95, Paper $19.95.

CASSAM, QUASSIM: *Self and World.* New York: Oxford, 1997. 208 p. Cloth $49.95.

CATES, DIANA FRITZ: *Choosing to Feel: Virtue, Friendship, and Compassion for Friends.* Notre Dame: University Press, 1997. x, 298 p. Cloth $32.00.

CLARK, STEPHEN R.L.: *Animals and Their Moral Standing.* New York: Routledge, 1997. viii, 194 p. Cloth $59.95, Paper $17.95.

COCKBURN, DAVID: *Other Times: Philosophical Perspectives on Past, Present, and Future.* New York: Cambridge, 1997. xiv, 355 p. Cloth $59.95.

COLES, ROMAND: *Rethinking Generosity: Critical Theory and the Politics of Caritas.* Ithaca: Cornell, 1997. xii, 246 p. Cloth $39.95, paper $16.95.

CONWAY, DANIEL W.: *Nietzsche and the Political.* New York: Routledge, 1997. xii, 163 p. Cloth $65.00, paper $16.95; and *Nietzsche's Dangerous Game.* New York: Cambridge, 1997. xii, 267 p. Cloth $54.95.

COUVALIS, GEORGE: *The Philosophy of Science: Science and Objectivity.* Thousand Oaks, CA: Sage, 1997. ix, 206 p. Cloth $69.95, paper $23.95.

CROMER, ALAN: *Connected Knowledge: Science, Philosophy, and Education.* New York: Oxford, 1997. xiii, 221 p. Cloth $25.00.

CUPIT, GEOFFREY: *Justice as Fittingness.* New York: Oxford, 1997. x, 185 p. Cloth $49.95.

CUTROFELLO, ANDREW: *Imagining Otherwise: Metapsychology and the Analytic Tradition.* Evanston, IL: Northwestern, 1997. xi, 166 p. Cloth $79.95, paper $22.50.

DAVIS, COLIN: *Levinas: An Introduction.* Notre Dame: University Press, 1997. viii, 168 p. Paper $17.00.

DECEW, JUDITH: *In Pursuit of Privacy: Law, Ethics, and the Rise of Technology.* Ithaca: Cornell, 1997. xii, 199 p. Cloth $39.95, paper $15.95.

DE GEORGE, RICHARD T.: *Academic Freedom and Tenure: Ethical Issues.* Lanham, MD: Rowman and Littlefield, 1997. xiii, 231 p. Cloth $62.50, paper $23.95.

DUNNING, STEPHEN N.: *Dialectical Readings: Three Types of Interpretation.* University Park: Penn State UP, 1997. vii, 191 p. Cloth $45.00, paper $19.95.

EBBS, GARY: *Rule-Following and Realism.* Cambridge: Harvard, 1997. xvi, 368 p. Cloth $42.50.

EDWARDS, JAMES C.: *The Plain Sense of Things: The Fate of Religion in an Age of Normal Nihilism.* University Park: Penn State UP, 1997. xi, 247 p. Cloth $45.00, paper $17.95.

ELFSTROM, GERARD: *New Challenges for Political Philosophy.* New York: St. Martin's, 1997. vi, 211 p. Cloth $45.00.

ELGIN, CATHERINE Z.: *Considered Judgment.* Princeton: University Press, 1997. x, 227 p. Cloth $32.50.

ERWIN, EDWARD: *Philosophy and Psychotherapy.* Thousand Oaks, CA: Sage, 1997. xi, 179 p. Paper $23.95.

EVERSON, STEPHEN: *Aristotle on Perception.* New York: Oxford, 1997. x, 309 p. Cloth $65.00.

FELDMAN, FRED: *Utilitarianism, Hedonism, and Desert.* New York: Cambridge, 1997. ix, 220 p. Cloth $54.95, paper $17.95.

FLEMING, MARIE: *Emancipation and Illusion: Rationality and Gender in Habermas's Theory of Modernity.* University Park: Penn State UP, 1997. vii, 243 p. Cloth $40.00, paper $17.95.

FLYNN, THOMAS R.: *Sartre, Foucault, and Historical Reason: Toward an Existentialist Theory of History, Volume One.* Chicago: University Press, 1997. xvi, 340 p. Cloth $55.00, paper $18.95.

FRAWLEY, WILLIAM: *Vygotsky and Cognitive Science: Language and the Unification of the Social and Computational Mind.* Cambridge: Harvard, 1997. 333 p. Cloth $39.95.

GALISON, PETER: *Image and Logic: A Material Culture of Microphysics.* Chicago: University Press, 1997. xxv, 955 p. Cloth $90.00, paper $34.95.

GANS, ERIC: *Signs of Paradox: Irony, Resentment, and Other Mimetic Structures.* Stanford: University Press, 1997. 222 p. Cloth $45.00.

GENDLIN, EUGENE F.: *Experiencing and the Creation of Meaning: A Philosophical and Psychological Approach to the Subjective.* Evanston, IL: Northwestern, 1997. xxix, 302 p. Paper $17.95.

GIUNTI, MARCO: *Computation, Dynamics, and Cognition.* New York: Oxford, 1997. xii, 177 p. Cloth $39.95.

GRAHAM, GORDON: *The Shape of the Past: A Philosophical Approach to History.* New York: Oxford, 1997. 236 p. Paper $14.95.

GRANT, RUTH W.: *Hypocrisy and Integrity: Machiavelli, Rousseau, and the Ethics of Politics.* Chicago: University Press, 1997. xiii, 201 p. Paper $22.50.

GRIFFITHS, PAUL E.: *What Emotions Really Are.* Chicago: University Press, 1997. 286 p. Cloth $27.50.

GROSS, MICHAEL L.: *Ethics and Activism: The Theory and Practice of Political Morality.* New York: Cambridge, 1997. xiv, 305 p. Cloth $59.95.

GYEKYE, KWAME: *Tradition and Modernity: Philosophical Reflections on the African Experience.* New York: Oxford, 1997. xv, 338 p. Cloth $45.00, paper $19.95.

HARRIS, GEORGE W.: *Dignity and Vulnerability: Strength and Quality of Character.* Berkeley: California UP, 1997. x, 148 p. Cloth $35.00.

HASELAGER, W.F.G.: *Cognitive Science and Folk Psychology: The Right Frame of Mind.* Thousand Oaks, CA: Sage, 1997. x, 165 p. Cloth $69.95, paper $23.95.

HAUERWAS, STANLEY, and CHARLES PINCHES: *Christians among the Virtues: Theological Conversations with Ancient and Modern Ethics.* Notre Dame:

University Press, 1997. xvii, 230 p. Cloth $29.95, paper $16.95.

HAUSMAN, CARL R.: *Charles S. Peirce's Evolutionary Philosophy.* New York: Cambridge, 1997. xvii, 230 p. Paper $18.95.

HERSH, REUBEN: *What Is Mathematics, Really?* New York: Oxford, 1997. xxiv, 343 p. Cloth $35.00.

HILL, CLAIRE ORTIZ: *Rethinking Identity and Metaphysics: On the Foundations of Analytic Philosophy.* New Haven: Yale, 1997. xviii, 180 p. Cloth $22.50.

HOLMES, ARTHUR F.: *Fact, Value, and God.* Grand Rapids, MI: Eerdmans, 1997. viii, 183 p. Paper $18.00.

HORNSBY, JENNIFER: *Simple Mindedness: In Defense of Naive Naturalism in the Philosophy of Mind.* Cambridge: Harvard, 1997. ix, 265 p. Cloth $35.00.

JENKS, CHRIS: *Childhood.* New York: Routledge, 1997. viii, 146 p. Cloth $55.00, paper $16.95.

KAGAN, SHELLY: *Normative Ethics.* Boulder, CO: Westview, 1997. 337 p. Cloth $69.00, paper $24.00.

KAHN, PAUL W.: *The Reign of Law: Marbury v. Madison and the Construction of America.* New Haven: Yale, 1997. xi, 306 p. Cloth $35.00.

KAUFMAN-OSBORN, TIMOTHY V.: *Creatures of Prometheus: Gender and the Politics of Technology.* Lanham, MD: Rowman and Littlefield, 1997. xi, 292 p. Cloth $57.50, paper $22.95.

KEGLEY, JACQUELYN ANN K.: *Genuine Individuals and Genuine Communities: A Roycean Public Philosophy.* Nashville: Vanderbilt, 1997. x, 260 p. Cloth $28.95

KEMAL, SALIM: *Kant's Aesthetic Theory: An Introduction, Second Edition.* New York: St. Martin's, 1997. xi, 197 p. Paper $17.95.

KIRBY-SMITH, H.T.: *A Philosophical Novelist: George Santayana and* The Last Puritan. Carbondale: Southern Illinois UP, 1997. xii, 208 p. Cloth $39.95.

KIVY, PETER: *Philosophies of Arts: An Essay in Differences.* New York: Cambridge, 1997. xi, 242 p. Cloth $59.95, paper $18.95.

KLEE, ROBERT: *Introduction to the Philosophy of Science: Cutting Nature at Its Seams.* New York: Oxford, 1997. xi, 258 p. Cloth $45.00, paper $23.95.

KLEIN, WAYNE: *Nietzsche and the Promise of Philosophy.* Albany: SUNY, 1997. xviii, 256 p. Paper $19.95.

KRAMER, MATTHEW H.: *Hobbes and the Paradoxes of Political Origins.* New York: St. Martin's, 1997. xii, 144 p. Cloth $65.00; and *John Locke and the Origins of Private Property.* New York: Cambridge, 1997. xiii, 347 p. Cloth $59.95.

KRETZMANN, NORMAN: *The Metaphysics of Theism: Aquinas's Natural Theology in* Summa Contra Gentiles I. New York: Oxford, 1997. xii, 302 p. Cloth $45.00.

LAD, FRANK: *Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction.* New York: Wiley, 1996. xix, 484 p. Cloth $74.95.

LEHRER, KEITH: *Self-Trust: A Study of Reason, Knowledge, and Autonomy.* New York: Oxford, 1997. ix, 204 p. Cloth $58.00, Paper $14.95.

LEMONCHECK, LINDA, and MANE HAJDIN: *Sexual Harassment: A Debate.* Lanham, MD: Rowman and Littlefield, 1997. x, 239 p. Cloth $52.50, paper $16.95.

LEPLIN, JARRETT: *A Novel Defense of Scientific Realism.* New York: Oxford, 1997. xii, 204 p. Cloth $39.95.

LESLIE, JOHN: *The End of the World.* New York: Routledge, 1996. vii, 310 p. Cloth $23.00

LEVI, ISAAC: *The Covenant of Reason: Rationality and the Commitments of Thought.* New York: Cambridge, 1997. xiv, 258 p. Cloth $59.95, paper $18.95.

LØGSTRUP, KNUD EJLER: *The Ethical Demand.* Notre Dame: University Press, 1997. xxxviii, 300 p. Paper $22.95.

MAHON, JOSEPH: *Existentialism, Feminism, and Simone de Beauvoir.* New York: St. Martin's, 1997. xi, 241 p. Cloth $65.00.

MANIN, BERNARD: *The Principles of Representative Government.* New York: Cambridge, 1997. ix, 243 p. Cloth $54.95, Paper $16.95.

MARCONI, DIEGO: *Lexical Competence.* Cambridge: MIT, 1997. xiii, 206 p. Cloth $25.00.

MARENBON, JOHN: *The Philosophy of Peter Abelard.* New York: Cambridge, 1997. xx, 373 p. Cloth $59.95.

MARTIN, C.J.F.: *An Introduction to Medieval Philosophy.* Edinburgh: University Press, 1997. viii, 148 p. Paper $25.00.

MASON, RICHARD: *The God of Spinoza: A Philosophical Study.* New York: Cambridge, 1997. xiv, 272 p. Cloth: $54.95.

MAY, TODD: *Reconsidering Difference: Nancy, Derrida, Levinas, and Deleuze.* University Park: Penn State UP, 1997. x, 208 p. Cloth $32.50, paper $16.95.

MCCARTHY, GEORGE E.: *Romancing Antiquity: German Critique of the Enlightenment from Weber to Habermas.* Lanham, MD: Rowman and Littlefield, 1997. xxvii, 378 p. Cloth: $67.50, paper $24.95.

MCCORMICK, JOHN P.: *Carl Schmitt's Critique of Liberalism: Against Politics as Technology.* New York: Cambridge, 1997. xii, 352 p. Cloth $39.95.

MCGINN, COLIN: *The Character of Mind.* New York: Oxford, 1997. vii, 176 p. Cloth $65.00, Paper $15.95; and *Ethics, Evil, and Fiction.* New York: Oxford, 1997. x, 186 p. Cloth $24.95; and *Minds and Bodies: Philosophers and Their Ideas.* New York: Oxford, 1997. viii, 264 p. Cloth $35.00.

MIDGLEY, MARY: *Utopias, Dolphins, and Computers: Problems of Philosophical Plumbing.* New York: Routledge, 1997. x, 182 p. Cloth $22.95.

MILLGRAM, ELIJAH: *Practical Induction.* Cambridge: Harvard, 1997. vi, 184 p. Cloth $29.95.

MILLS, CHARLES W.: *The Racial Contract.* Ithaca: Cornell, 1997. xii, 171 p. Paper $19.95.

MOHANTY, J.N.: *Phenomenology: Between Essentialism and Transcendental Philosophy.* Evanston, IL: Northwestern, 1997. xii, 105 p. Cloth $59.95, paper $19.95.

MOODY-ADAMS, MICHELE M.: *Fieldworks in Familiar Places: Morality, Culture, and Philosophy.* Cambridge: Harvard, 1997. x, 260 p. Cloth $35.00.

MOORE, A.W.: *Points of View.* New York: Oxford, 1997. xiv, 313 p. Cloth $35.00.

MORRISON, ROBERT G.: *Nietzsche and Buddhism: A Study in Nihilism and Ironic Affinities.* New York: Oxford, 1997. x, 250 p. Cloth $62.00.

MOUNCE, H.O.: *The Two Pragmatisms: From Peirce to Rorty.* New York: Routledge, 1997. viii, 245 p. Cloth $69.95, paper $18.95.

MULLER-SIEVERS, HELMUT: *Self-Generation: Biology, Philosophy, and Literature around 1800.* Stanford: University Press, 1997. 222 p. Cloth $39.50.

NAGEL, THOMAS: *The Last Word.* New York: Oxford, 1997. viii, 147 p. Cloth $19.95.

NORRIS, CHRISTOPHER: *Resources of Realism: Prospects for 'Post-Analytic' Philosophy.* New York: St. Martin's, 1997. xviii, 247 p. Cloth $65.00; and *Against Relativism: Philosophy of Science, Deconstruction, and Critical Theory.* Malden, MA: Blackwell, 1997. xi, 330 p. Paper $24.95.

NOZICK, ROBERT: *Socratic Puzzles.* Cambridge: Harvard, 1997. viii, 400 p. Cloth $35.00.

O'CONNELL, ROBERT J.: *Plato on the Human Paradox.* New York: Fordham, 1997. xvii, 162 p. Cloth $30.00, Paper $15.00.

O'DONOVAN-ANDERSON, MICHAEL: *Content and Comportment: On Embodiment and the Epistemic Availability of the World.* Lanham, MD: Rowman and Littlefield, 1997. viii, 167 p. Cloth $53.00, paper $21.95.

OLIVER, KELLY: *Family Values: Subjects between Nature and Culture.* New York: Routledge, 1997. xviii, 259 p. Paper $16.95, Cloth $59.95.

O'NEILL, SHANE: *Impartiality in Context: Grounding Justice in a Pluralist World.* Albany: SUNY, 1997. x, 288 p. Paper $18.95.

PASNAU, ROBERT: *Theories of Cognition in the Later Middle Ages.* New York: Cambridge, 1997. xi, 330 p. Cloth $54.95.

PEIRCE, CHARLES SANDERS: *Pragmatism as a Principle and Method of Right Training: The 1903 Harvard Lectures on Pragmatism.* PATRICIA ANN TURRISI, ed. Albany: SUNY, 1997. xi, 305 p. Paper $18.95.

PEPERZAK, ADRIAAN THEODOOR: *Beyond: The Philosophy of Emmanuel Levinas.* Evanston, IL: Northwestern, 1997. xviii, 248 p. Cloth $79.95, paper $22.50.

PETTIT, PHILIP: *Republicanism: A Theory of Freedom and Government.* New York: Oxford, 1997. x, 304 p. Cloth $29.95.

PIPPIN, ROBERT B.: *Idealism as Modernism: Hegelian Variations.* New York: Cambridge, 1997. xiii, 466 p. Cloth $59.95, paper $19.95.

PLATTS, MARK: *Ways of Meaning: An Introduction to a Philosophy of Language.* Cambridge: MIT, 1997. xiv, 304 p. Paper $25.00.

POSNER, RICHARD H.: *Law and Legal Theory in England and America.* New York: Oxford, 1997. x, 134 p. Cloth $24.95.

POSTER, MARK: *Cultural History and Postmodernity: Disciplinary Readings and Challenges.* New York: Columbia, 1997. ix, 173 p. Cloth $47.50, paper $16.50.

PRESTON, JOHN: *Feyerabend: Philosophy, Science and Society.* Malden, MA: Blackwell, 1997. xiv, 234 p. Paper $21.95.

REINEKE, MARTHA J.: *Sacrificed Lives: Kristeva on Women and Violence.* Bloomington: Indiana UP, 1997. x, 232 p. Cloth $35.00, paper $18.95.

THE JOURNAL OF PHILOSOPHY

VOLUME XCV, NO. 5, MAY 1998

WHY IS A WING LIKE A SPOON? A PLURALIST THEORY OF FUNCTION*

It is commonplace among function theorists to assume that a viable function theory will be equally applicable to biological traits, artifacts, and anything else to which functions can be ascribed. For example, Ruth Garrett Millikan[1] speculates that biological traits, language devices, and tools are alike in three very important respects. They all have functions that they sometimes fail to serve; they all have forms that are sometimes (in the case of biological traits and tools) or always (in the case of language devices) arbitrary in relation to their functions; and they all have *proper functions*—that is, there is something specific they are supposed to do, even though they may never perform this function, or may be temporarily coopted for some other use. The one difference she notes is that the proper functions of biological devices, language devices, and representational mental states are not the result of purposive design, as are the proper functions of tools. But, clearly, the similarities outweigh this single dissimilarity.

The nature of artifacts generally, and the nature of their functionality in particular, is taken to be so transparently obvious, however, that virtually nobody has bothered to examine it at any length. Discussion of tools and other artifacts in the texts of function theorists consists of scattered examples and isolated paragraphs. These authors are also not relying on treatises or analyses from the wider philosophical literature; for there are virtually none to cite in the en-

---

* The final version of this paper was improved by suggestions from Charles Cross, Robert Cummins, Ruth Millikan, John Post, and Mark Risjord.

[1] *Language, Thought, and Other Biological Categories: New Foundations for Realism* (Cambridge: MIT, 1984), introduction

0022-362X/98/9505/215–54

tire history of Western philosophy.[1] So we have a rather curious situation. On the one hand, it is altogether common to assume that a theory of artifact function will simply fall out of the general theory of function now being articulated, even though this purportedly general theory is based on an examination of biological function. On the other hand, the realm of artifactual function has received only the most cursory scrutiny in its own right. So it appears to be an article of faith among function theorists that (a) an adequate general theory of function can be arrived at without adverting to the functionality of artifacts, and (b) once that general theory is in place, a few minor adjustments will serve to fit it to the artifactual case.

I think both parts of this article of faith are mistaken. I shall address primarily part (a), although some comments about part (b) are contained in section III. Specifically, I contend that only a pluralist theory comprising two distinct notions of what it is to be a function will serve as an adequate general theory. This point can be glimpsed, but only just, from the vantage point of biological function. It is when you consider artifactual function that it becomes patently obvious only a pluralist theory will do.

Section I of this paper lays out the two notions of function comprising the pluralist theory. Section II shows why both notions are necessary, by way of showing that concerted attempts to do away with one of them fail. This demonstration draws on examples from the artifactual realm to motivate major points of the argument. Section III is an outline of artifact function. It confirms the conclusions of section II, and also begins the task of describing some of the special features of artifactual function which will need accommodation within the general theory of function in the form of revisions or extensions.

---

[1] In a recent and path-breaking work on what he calls artifact theory, Randall Dipert notes:

> What is missing [in the history of philosophy] is any sustained, developed account of artifacts, or even an attempt to give one. There is no major treatise on artifacts, no major section of Locke's *Treatise* or one of Kant's *Critiques* that deals squarely with artifacts. There is not even a major recent, general article on the subject—*Artifacts, Art Works, and Agency* (Philadelphia: Temple, 1993), p. 8.

This appraisal might be qualified by noting that there is now a considerable literature in philosophy of technology. But although technology is usually defined in this literature so as to comprise all the commonplace implements and artifacts of daily life, these are again not the focus of sustained examination. This is because philosophers of technology are primarily concerned with the changes wrought on human beings and their societies by the sophisticated recent technologies of the industrial West. But in any case, function theorists do not cite the philosophy of technology literature in support of their claims about tools and other artifacts.

I. SYSTEM FUNCTION AND PROPER FUNCTION

The concept of function was of interest in the first instance to philosophers of science trying to understand the ascription of function to biological organs, such as the heart. Subsequently, philosophers of mind acquired an interest in function, motivated by the hypothesis that to ascribe meaning to a mental representation is, at least in part, to say what its function is in the mental economy of the organism. But consensus on a theory of function has proved elusive. The main problem, as I shall show, is that there are two distinct notions of function being articulated in the literature. The difficulties arise either from a failure to distinguish them adequately, or from a mistaken insistence that one or the other of them is *the* correct notion. In fact, as we shall see, the two notions are not competitors, but are complementary, and make distinct contributions to a general theory of function. In this section, I concentrate on a few well-known and representative accounts that will bring these two notions of function into focus. The following sections will lay out the complementary roles they play in a pluralist theory of function incorporating both of them on equal terms.

In a pivotal article, Larry Wright[3] set out to give an analysis of function that would avoid the obvious difficulties with some earlier analyses. It has two parts:

> The function of *X* is *Z* *means*
> (a) *X* is there because it does *Z*,
> (b) *Z* is a consequence (or result) of *X*'s being there.
>
> The first part, (a), displays the etiological form of functional ascription-explanations, and the second part, (b), describes the convolution which distinguishes functional etiologies from the rest (*ibid.*, p. 161).

It is part (a) that is of interest for my purposes. This is itself a two-part criterion for the ascription of function. One part seems etymologically straightforward: the function of something has to do with what that something *does*, some performance in which it is involved.[4] This performance may be either active (for example, the heart circulates blood) or passive (for example, the chair supports seated humans). More importantly, since at any given moment the thing may not be executing the relevant performance, this criterion turns out to require a specification of current capacities or dispositions to perform under appropriate circumstances. For example, the heart has the capacity/disposition to circulate blood, provided the lungs are supplying the heart muscle with oxygen, the nervous system is relaying the necessary signals, and so on. These conditions may be in abeyance during open-heart surgery, for example, but the capacity/disposition must still be there, ready to kick in when the right conditions are reinstated, in order for circulating blood to be the heart's ongoing function.

---

[3] "Functions," *Philosophical Review*, LXXXII (April 1973): 139-68.

[4] 'Function' is from Latin *fungi*, to perform.

But a thing may engage in all sorts of performances, many of which have nothing to do with its function, taken strictly. For example, the heart both circulates blood and makes a characteristic noise. Wright insists on an analysis that will draw what he calls the *function-accident* distinction in the right way, clearly separating performances which are a matter of fulfilling a function from those which are adventitious. In the case of the heart, it must rule out noise making and rule in circulating blood. His analysis does this by connecting function ascriptions to performances that account for why the thing is *there* in the first place. It is important to note that 'there' is equivocal in Wright's formulation, since it can mean either why something is in the location it is in, or why something is in existence at all. In either case, though, what you have to come up with is a historical account. In the case of biological organs like the heart, this will be its evolutionary history as mediated by natural selection. In the case of artifacts like the chair, it will be a history of deliberate selection by a user or community of users. Thus, the heart is standard issue for vertebrates now, not because it makes noise, but because it circulates blood. Similarly, chairs continue to be made and purchased, not because you can stand on them to get things off high shelves, but because they support humans in a sitting position. The underlying point here is that the nature of functional explanation is fundamentally causal.

> Putting the matter in this way suggests that functional ascription-explanations are in some sense etiological, concern the causal background of the phenomenon under consideration. And this is indeed what I wish to argue: functional explanations, although plainly not causal in the usual, restricted sense, do concern how the thing with the function *got there*. Hence they *are* etiological, which is to say "causal" in an extended sense (*ibid.*, p. 156).

So, for Wright, there are two things you have to do to produce a function ascription in the service of a functional explanation. First, you have to look at what the thing in question has the *capacity/disposition* to do under appropriate circumstances. Second, you have to look at the *history* of the thing in question so as to determine which of its capacities/dispositions account for its being there.

The problem with Wright's analysis is that the two criteria—selection history and current capacity/disposition—pull in opposite directions and have proved impossible to weld together into a harmonious single account of functional ascription-explanation. An early and significant challenge to Wright's position along these lines came from Robert Cummins,[5] who pointed out that the current performance of a biological organ does not necessarily have anything to do with the evolutionary reasons why it is there. In this passage, he discusses the contractile vacuoles possessed by protozoans, which in fresh-water species perform the crucial task of expelling excess water from the organism.

> For it might seem that natural selection provides the missing causal link between what something does in a certain type of organism and its presence in that type of organism. By performing their respective functions, the contractile vacuole and the neurofibrils help species incorporating them [protozoans and ciliates, respectively] to survive, and thereby contribute to their own continued presence in organisms of those species, and this might seem to explain the presence of those structures in the organisms incorporating them.
>
> Plausible as this sounds, it involves a subtle yet fundamental misunderstanding of evolutionary theory. A clue to the mistake is found in the fact that the contractile vacuole occurs in marine protozoans that have no excess-water problem but the reverse problem. Thus the function and effect on survival of this structure is not the same in all protozoans. Yet the explanation of its presence in marine and fresh-water species is almost certainly the same. This fact reminds us that the processes actually responsible for the occurrence of contractile vacuoles in protozoans are totally insensitive to what that structure does (*ibid.*, p. 750).

Other biological examples can be adduced, such as the panda's "thumb" (actually a modified wristbone) made famous by Stephen Jay Gould,[6] or the wings of flightless birds, which may be used (for example, by penguins) for swimming. But this phenomenon is more common in the realm of artifacts, where things designed for one function frequently end up being used at least as often for something else. The case of the chair routinely used to stand on in order to get things off high shelves has already been mentioned. Or consider the old tire hung from a tree branch which is now a swing; the jelly jar that is now a drinking glass; and the orange crate that is now a bookcase or a makeshift living-room coffee table. In such cases, there is indeed a story about why these items are *located* where they are which does appeal to their current capacities/dispositions—it is because the orange crate has a nice flat top that it makes a good coffee table. But there is a completely different story to be told about why these things are there in the sense of being in existence at all, and it need make no such appeal since the thing may no longer even have the capacity/disposition that is responsible for its coming into existence. The tire that is now a swing is almost certain to be treadless, and the penguin's wings are not capable of lifting it off the ground. Such cases show that *current* capacity/disposition and selection history are in principle divergent. The difficulty, in short, is that Wright's formulation '*X* is there because it does *Y*' does not necessarily hold. *X* may be there because it *used* to do *Y*, although now it cannot do *Y* at all, as in the case of the tire swing. Or it may be there because its ancestors *did Y*, although it has never been able to do *Y* itself, as in the case of the penguin's wings.

---

[5] "Functional Explanation," this JOURNAL, LXXII, 20 (November 20, 1975): 741-64.

[6] "The Panda's Thumb," in *The Panda's Thumb: More Reflections in Natural History* (New York: Norton, 1980), pp. 19-26

The underlying problem, according to Cummins, is that Wright and other previous writers on this topic misunderstood how functional explanation in science works. It is not, as they assume, a causal-historical account of how something got there. Rather, it is an account of the current capacities of a whole system in terms of the capacities of its components, just as the capacity of a Detroit assembly line to produce a car can be explained in terms of the tasks accomplished by the workers at each station.

> Functional analysis in biology is essentially similar. The biologically significant capacities of an entire organism are explained by analyzing the organism into a number of "systems"—the circulatory system, the digestive system, the nervous system, etc.—each of which has its characteristic capacities. These capacities are in turn analyzed into capacities of component organs and structures. Ideally, this strategy is pressed until pure physiology takes over... (*op. cit.*, pp. 760-61).

The analyzing capacities of the components of the system are, according to Cummins, precisely what we want to call the functions of those components—the functional roles they play within the system as a whole. So in this case, functional ascription-explanation does have a necessary connection to the current capacities/dispositions of things, to what they do. Cummins thus retains the second criterion from Wright's analysis of function, but replaces the first criterion—the one that claims that functional explanation is based on selection history—with an alternative notion of functional explanation which appeals only to current system context. In what follows, I shall refer to this notion of function ascription-explanation as *system function*.[7]

Even if the justice of Cummins's criticism of analyses like Wright's is granted, however, the plausibility of his alternative analysis as a definitive and complete account of functional ascription-explanation is weak, since he fails to respond adequately to the consideration motivating Wright's claim that functional explanation must appeal to selection history. What drove Wright to this conclusion was the necessity of separating out from all the transient and accidental performances in which things may engage those which are tied in with their function, strictly speaking. Intuitively, for Wright, that the heart makes noise is accidental, a mere by-product of its "real" function, which is pumping blood. Adverting to the selection history of hearts confirms this intuition. Cummins's account, on the other hand, makes it impossible to draw the function-accident distinction where Wright wants to draw it. This is because the notion of a system is extremely open-ended, comprising almost any orderly arrangement of interacting things. Thus, for the majority of performances something manifests, it is possible to describe a system in which it is a functioning component. So, for example, if you consider the doctor-patient relationship as a diagnostic system, the sounds the heart makes do indeed appear as a functional component, interrelated with others, such as the way various organs look or feel to the touch, the results of various tests and measurements, and so on. Cummins tries to downplay this striking inclusiveness of his notion of system function by pointing out that, in some cases, it will not be possible to specify any sort of interesting containing system for a given performance, so that there will indeed be cases where system-functional explanation is just not possible or appropriate. For example, Wright mentions the case of a belt buckle saving its wearer's life by deflecting a stray bullet. Here, it is hard to imagine what system could be described such that this singular event would come out as a function of the buckle rather than as just a lucky accident. So a Cummins-style analysis and a Wright-style analysis would draw the function-accident distinction in the same place in cases like this.

---

[7] Cummins functions are called *causal role functions* by Karen Neander—"Functions as Selected Effects: The Conceptual Analyst's Defense," *Philosophy of Science*, LVIII (1991): 168-84. I prefer *system function* because not all such functions correspond straightforwardly to an integral causal role. For example, one of the functions of the queen in the chess system is to capture pawns. But her capacity to do this does not cause pawns to be captured in the straightforward way that the capacity of the heart to pump blood causes blood to be circulated through the body. Moreover, in *The Nature of Psychological Explanation* (Cambridge: MIT, 1983), Cummins himself explicitly contrasts functional analysis (or interpretive explanation, as he also calls it) with analysis in terms of causal laws.

But Cummins's account still draws this line a great deal more liberally than Wright and many others are willing to countenance. It is, therefore, no surprise to find Millikan seconding Wright's contention that functional explanation is causal and that, in consequence, selection history is essential to it. On this foundational assumption, she constructs an account of what she calls *proper function.*[8]

> Biological categories[9] are carved out not by looking at the actual structure, actual dispositions, or actual functions of the organ or system that falls within the category, but by looking at (or speculating about) its history.... In like manner, every body organ or system falls in the biological or physiological categories it does due to its historical connections with prior examples of kinds that have served certain functions or, typically, sets of functions. So whether or not it is itself capable of serving any of these functions, every organ or system is associated with a set of functions that are biologically "proper" to it, functions that have helped account for the survival and proliferation of its ancestors. I call these functions "proper functions" of the organ or system.[10]

Hearts, for example, have proliferated not because they make noise, but because they circulate blood. Thus, even a heart that comes into the world so badly deformed that it is never capable of circulating a single drop still has circulating blood as its proper function. It is still relevantly similar to its ancestors which did circulate blood, and whose circulation of blood in the past accounts for the production of the deformed heart (though not for its deformation) in the present. Similarly, vegetable peelers have proliferated—this time in virtue of deliberate selection on the part of manufacturers and consumers—because they remove peels from fruits and vegetables. So even if a manufacturing defect makes it impossible for one of them to peel anything at all, peeling is still its proper function.

---

[8] This account appeared in *Language, Thought, and Other Biological Categories.* Later, Millikan said that she had proposed it without having any knowledge of Wright's previous work—*White Queen Psychology and Other Essays for Alice* (Cambridge: MIT, 1993), p. 33. Nor was she familiar with Cummins's work at that time (Millikan, personal communication). So she was not intending to contribute to the Wright-Cummins debate about function. Yet her theory of proper function does, in fact, constitute a major advance in that debate, and has been treated as such in the literature, as well as by Millikan herself in subsequent writings.

[9] For Millikan, anything that has a selection history of any kind, and can therefore be analyzed in accordance with the definition of proper function, is a member of a biological category. Thus, language, thought, and artifacts are all biological categories.

[10] *White Queen Psychology*, pp. 55-56.

An analysis in accordance with Millikan's notion of proper function draws the function-accident line where Wright thinks it should be drawn, and it does so in virtue of the same reliance on selection history. It is not any old performance a thing engages in which constitutes its proper function, but only those specific performances which enabled its ancestors to survive and reproduce. On the other hand, Millikan agrees with Cummins's contention that selection history and current capacity/disposition do not necessarily march in lockstep, as Wright's analysis assumes, but may in fact march in quite different directions. So instead of connecting selection history directly with what a thing now does—its current capacity/disposition—she connects selection history with what the ancestors (or prototypes, in the case of artifacts) of a thing did in the past. So where Wright says '*X* is there because *it does Y*', Millikan says '*X* is there because some of its *ancestors did Y*'. This fixes the problem noted by Cummins, since what the precursors of a thing did just is part of the selection history that accounts for its existence, and the problematic current capacity/disposition is not appealed to at all.

A distinguishing feature of Millikan's notion of proper function is that a correlative notion of malfunction falls out directly. If you can say what a thing is supposed to do, then you can also say when it is failing to do something it is supposed to do, that is, malfunctioning. This is of crucial importance to Millikan because she has her sights set on thought and language. Specifically, she wants a function-based theory of intentionality and representation, and this requires a correlative theory of failed intentionality and misrepresentation. Lack of such a correlative theory has long bedeviled function-based theories of intentionality in philosophy of mind, because they have almost without exception relied implicitly on some variation of the notion of system function, which in turn relies on current capacity/disposition as the basis for function ascription.[11]

The problem is that being locked into an appeal to current capacity/disposition for the determination of function—as both Wright and Cummins are in their different ways—makes it difficult to distinguish between malfunction and just not having a function. Consider someone born blind. Her eyes simply never have the capacity to see, and so on Cummins's analysis, where the sensory system is the containing system, they do not contribute to the functioning of the whole sensory system at all, and so do not have a function. Wright's analysis yields a similar result, because you cannot complete the formula '*X* is there because it does *Y*' in any reasonable way. The eyes of a congenitally blind person are just not there because of any current capacity/disposition they have, regardless of whether 'there' is construed as mere location or as sheer existence. So it seems they do not have a function on this analysis either. Millikan's point is that it would be useful for us to have a notion of function that would enable us to say that they do have a function, but are unable to perform it, that is, that they are malfunctioning components of the sensory system. It is only when we let go of current capacity/disposition as a determining criterion of function that we are able to do this.

---

[11] For more on this issue, see Fred Dretske, "Misrepresentation," in Radu Bogdan, ed., *Belief: Form, Content, and Function* (New York: Oxford, 1986), pp. 17-36; Jerry Fodor, *Psychosemantics: The Problem of Meaning in the Philosophy of Mind* (Cambridge: MIT, 1987); and Millikan, *White Queen Psychology*, especially the introduction and chapters 4 and 6.

Thus, in addition to drawing the function-accident distinction very liberally, the notion of system function is also resolutely nonjudgmental. Things either have a capacity/disposition or they do not; but on this view, there is no particular sense to be made of the claim that there are particular capacities/dispositions they *ought* to have, but are temporarily or permanently unable to exercise even under conditions that seem completely appropriate or normal. The main difference between system function and proper function is therefore that the latter is normative whereas the former is not. Moreover, the normativity that defines proper function is not reducible to statistical regularity. It might happen that even the vast majority of a certain type of thing are unable to fulfill their proper function. Millikan employs the telling example of sperm, most of which do not or cannot fertilize an ovum, although that is still what each of them is supposed to do. Millikan uses the capitalized term 'Normal' (hereafter 'normal') to signify this strict sense of normativity.[12]

This difference in normative status accounts for both further differences we have noted. Unlike proper function, system function has no strong commitment to a distinction between "real" function and adventitious functions a thing may have transiently or in addition; and it lacks machinery for specifying malfunctions in contradistinction to mere lack of function. Normative theories, on the other hand, are strongly committed to there being capacities/dispositions of things which they ought to have. So not having them is interpreted on this view as a positive defect; and any other capacities/dispositions they may have are regarded as irrelevant.

---

[12] For more elaboration of this point, see *White Queen Psychology*, especially chapter 3.

So Wright's original notion of function has split in two. Cummins took one element of Wright's analysis—the idea that function has to do with current capacity/disposition, jettisoned selection history as irrelevant to it, and came up with the nonnormative notion of system function. Millikan took the other element of Wright's analysis—the idea that function has to do with selection history, jettisoned current capacity/disposition as irrelevant, and came up with the normative notion of proper function. Each of these notions of function has associated with it a proprietary mode of explanation. System function is associated with explanations taking the form of compositional analyses of the capacities of containing systems in terms of their component parts. Proper function is associated with explanations taking the form of causal-historical accounts of why a thing is there in the first place.

But how are these two notions of function related to each other? Are they simply rivals, in the sense that one is *the* correct account of function and the other is just wrong? Or are they both correct and useful accounts, such that we simply have to settle for there not being a single, unitary account of function? In one place, Millikan[13] takes a pluralist view.

> In the context of biology, an obvious anchor for Cummins' notion of function is the system that is the life cycle of a species, a most salient complex capacity to be explained is the capacity of a given species to survive, develop to adulthood and reproduce. Now it is evident that any Cummins-style functional explanation of how a species has recently been surviving, developing and reproducing will assign functions to very many parts and traits of individual organisms that *coincide* with selected [that is, proper] functions of these parts and traits. The coincidence will not necessarily be complete, however. If there are traits that must be mentioned in explaining how the life cycle has recently been accomplished but that were not in fact ever selected for, then these will have Cummins [system] functions but not selected [proper] functions. And if there are traits that were once selected for but that have not recently been contributing to the life cycle, these will have selected [proper] functions but not Cummins [system] functions (*ibid.*, p. 175).

A tendency to conflate these two notions of function is thus attributable to the fact that in many—perhaps most—cases, either analysis will yield the same function specification. It is only the crucial and somewhat less common cases in which only one of these notions yields a function specification at all which clearly indicate the presence of two separate notions. This conflationary tendency is present in Wright's account of function, and explains why it eventually split apart along a principled line of cleavage, as described above.

---

[13] "An Ambiguity in the Notion 'Function'," *Biology and Philosophy*, IV (1989): 172-76.

Other function theorists have verified this split and followed Millikan's lead toward pluralism.[14] But this pluralistic attitude is far from an established consensus. It is still assumed by many writers that there should be *a* unitary conception of function, applicable to all cases. Under this assumption, if there is more than one conception of function floating, either one of them must be just wrong, or else there must be a deeper ground of unity we have failed to recognize. Philip Kitcher[15] takes the latter tack, arguing that proper function and system function can be unified under the rubric of design. This attempt at unification has been critically examined by Peter Godfrey-Smith (*op. cit.*), and I shall not take it up here. More importantly, though, Millikan herself often implicitly takes the former tack by insisting, in effect, that all interesting cases of functions are cases of proper function, thus relegating system function to the status of insignificance, if not entirely to oblivion. This is not consistent with her explicit endorsement of pluralism in the theory of function. Examining this inconsistency more closely will enable us to see why proper function and system function must be recognized not merely as distinct types of function, but as equally important for a viable general theory of function.

II. ADAPTATION AND EXAPTATION

Millikan's implicit rejection of pluralism in the theory of function is best diagnosed through her discussion of exaptation. *Exaptation* is the biologically defined form of what I have been calling system function, and is specifically distinguished from *adaptation*, the biologically defined form of proper function.

> Following Williams...we may designate as an *adaptation* any feature that promotes fitness and was built by selection for its current role (criterion of *historical genesis*). The operation of an adaptation is its *function*.... We may also follow Williams in labelling the operation of a useful character

---

[14] See, for example, Karen Neander; and Peter Godfrey-Smith, "Functions: Consensus without Unity," *Pacific Philosophical Quarterly*, LXXIV (September 1993): 196-208.

[15] "Function and Design," in P. A. French, T. E. Uehling, Jr., and H. K. Wettstein, eds., *Midwest Studies in Philosophy*, Volume XVIII (Notre Dame: University Press, 1993), pp. 379-97.

> not built by selection for its current role as an *effect*.... We suggest that such characters, evolved for other usages (or for no function at all), and later "coopted" for their current role, be called *exaptations*.... They are fit for their current role, hence *aptus*, but they were not designed for it, and are therefore not *ad aptus*, or pushed towards fitness. They owe their fitness to features present for other reasons, and are therefore *fit (aptus) by reason of (ex)* their form, or *ex aptus*.[16]

Adaptation here clearly corresponds to Millikanian proper function, as it depends on selection history. Millikan's understanding of selection is inspired by evolutionary biology. It is a matter of differential survival and reproduction.

> The functional trait must be one that is there in *contrast* to others that are *not* there, because of historic difference in the results of these alternative traits.... The difference between merely luckily or accidentally cycling through successive generations and having genuinely functional traits whose biological *purpose* is to contribute to such cycling lies in whether these traits were once selected from among other once extant traits due to their superior capacities to continue the cycle. Graphically, whether my shoulders have as a biological function to hold up my clothes depends not on what proportion of my ancestors used their shoulders that way to advantage but on whether there were once shoulderless people who died out because they had nothing to hang their clothes on.[17]

Clearly, the implication here is that shoulders do not have the proper function of holding up clothes, because they do not have the sort of selection history required. Similarly, the ubiquitous Panglossian question of whether noses have the proper function of supporting eyeglasses must be answered in the negative for the same reason. This selection-history criterion is a strong one, in the sense that a protracted period of time is required during which the performance of the trait is tested against alternatives and found successful, thus ensuring its own reproduction. In particular, Millikan says explicitly that the first few instances of the performance do not yet count as exercises of a proper function.

> To say that trait *t* has function *f* is to say that *t* had a history during which it was selected for doing *f*.... Of course, *f* was not the function of

---

[16] Gould and Elisabeth S. Vrba, "Exaptation: A Missing Term in the Science of Form," *Paleobiology*, VIII (1982): 4-15, here p. 6. See also Gould and Richard C. Lewontin, "The Spandrels of San Marco and the Panglossian Program," *Proceedings of the Royal Society of London*, CCV (1979): 281-88; and Lewontin, "Adaptation," in Elliott Sober, ed., *Conceptual Issues in Evolutionary Biology* (Cambridge: MIT, 1984), pp. 235-51.

[17] *White Queen Psychology*, p. 38.

> the very first *t* tokens selected, even though they did *f* and that is why they were selected. They were not selected "because they had functions" but because they did *f* (*ibid.*, pp. 40-41)

Although Gould and Elisabeth S. Vrba decline to use the term 'function' to characterize the operation of an exaptation, this notion corresponds to system function in Cummins's sense, since it is a useful performance based solely on the form of something, that is, on its current capacities and dispositions, regardless of how they got there. This point is borne out by the examples Gould and Vrba give of exaptation. For instance, they point to the common hypothesis that birds evolved feathers for insulation first, which were only later turned to account and further modified for flight. Similarly, some herons use their wings for "mantling," spreading the wings to cast a shadow so that they can see small fish in the water below. Another example frequently cited in this debate is the use aquatic turtles make of their hind flippers to dig holes laboriously in which to lay their eggs. (Terrestrial turtles do the same, equally laboriously, with their hind legs.) Since these limbs are *ad*apted for locomotion, this use is an exaptation. These examples display one important difference. The use of flippers or legs for digging and the mantling use of wings are transient cooptations, as the limb can be immediately returned to its original proper function. The flight feathers of birds, on the other hand, have been modified in order to suit their new function in ways that may compromise their original insulating function. But all of these examples nevertheless illustrate the point that some traits turn out to be useful for performing certain functions because of their current capacities or dispositions, even though the trait is not there in the first place because its ancestors were used for that purpose. Thus, an exaptation is a trait with a system function.

Millikan's pluralist tolerance of system function should translate into a similar tolerance of exaptation, but in fact she is very critical of this notion, arguing that the vast majority of the phenomena taken to be exaptations are actually adaptations, that is, that the functions in question are actually *proper* functions rather than *system* functions. To make this case, she refines the basic notion of selection she uses to define proper function. The first refinement is simply a reminder about what natural selection does in the biological realm, while the second is an expansion of her original definition of proper function. Together, these refinements have the effect of drawing into the realm of proper-functional adaptation virtually the whole range of phenomena alleged to fall under system-functional exaptation.

The reminder is that selection does not build traits or organs so much as it differentially maintains them. Of course, it does build complex organs like the eye by selecting for or against new ocular subtraits as they arise. But traits arise by other mechanisms, and selective forces then may actively maintain them in existence or may extinguish them, depending on whether they are beneficial or deleterious to the organism. This is important because it reminds us that something which starts out as a system-functional exaptation can rapidly become a proper-functional adaptation. For example, the hind limbs of turtles or the flight feathers of birds are now presumably being actively maintained by selective forces operating on their more recently acquired performances as excavators and airfoils respectively. There are two possible scenarios here. Either a second, unrelated proper function can be acquired to operate in parallel with the original one, or a change of proper function occurs such that the original proper function is no longer exercised or is lost entirely. But in either case, Millikan would claim, what you have in the long run is another proper-functional adaptation, not an ongoing system-functional exaptation. We shall return to this point at the end of this section.

The expansion of the original definition of proper function is, in my view, connected to something Millikan called *derived proper function* in her original account.[18] This is a connection which Millikan herself did not intend, and is not now disposed to accept.[19] I think it is essential to press the point, however, because the notion of derived proper function is, as I shall show, inconsistent with her basic conception of proper function in terms of selection history. This fundamental inconsistency is ultimately responsible for her inconsistent position with regard to pluralism in the theory of function.

The definition of proper function we have been working with so far is a definition of *direct proper function* only—functions of something that are specifiable directly in terms of the selection history of that thing itself. *Derived proper functions*, by contrast, are functions something gets indirectly in virtue of being produced by another thing that already has the relevant proper function.

> A originated as the product of some prior device that, given its circumstances, had performance of F as a proper function and that, under those circumstances, normally causes F to be performed by *means* of

---

[18] For this exposition, see *Language, Thought, and Other Biological Categories*, chapter 2.

[19] Millikan, personal communication.

> producing an item like A. Items that fall under [this] condition have "derived proper functions," functions derived from the function of the devices that produce them.[20]

Suppose you have a chameleon whose skin has turned a novel (for chameleons) shade of brown to match its surroundings. There is a mechanism in the skin of the chameleon which has the direct proper function of changing the chameleon's skin color to match its environment, and the further direct proper function of making the chameleon invisible to predators by so doing. But making the chameleon invisible to predators cannot be the *direct* proper function of that novel shade of brown, since it has not been specifically selected for. Indeed, by hypothesis, no chameleon has ever turned that particular shade of brown before. So Millikan says it has making the chameleon invisible to predators as a *derived* proper function—derived from the skin-coloration mechanism which produced it, and which does have making the chameleon invisible to predators as a direct proper function.

Another example is provided by language tokens.

> Language-device families, hence tokens of these, have (we speculate) *direct* proper functions too; but tokens of these have also *derived* proper functions—derived from the speaker's intentions. The direct and the derived proper functions of a language token often are not the same, for example, whenever a speaker does not understand how to use a word or uses it sarcastically or metaphorically.[21]

So, if I intend to instruct you to turn left, and I say 'turn right', my utterance has the *direct* proper function of meaning 'turn right' due to its selection history as a phrase in the English language; but the *derived* proper function—derived from the proper function of my intention—of meaning 'turn left'. On Millikan's view, the proper function of my intention is already a derived proper function—derived from the direct proper functions of the basic cognitive mechanisms that ultimately produced it. So, both in cases where something has no direct proper function of its own (the chameleon's novel shade of brown) or in cases where it does (language tokens), a thing can still derive a proper function in virtue of being produced by something else that has that proper function either directly or by derivation.

This derivation happens, Millikan says, "without the mediation of additional selection processes."[22] The underlying idea is that certain proper functions of a thing can be directly communicated to something else when there is a particular kind of production relation between them. What makes this idea plausible, I think, is the sense that the communication of the proper function *is* mediated by a sort of secondary selection. For example, the chameleon's skin-coloration mechanism, which has been selected in the primary, biological sense for making the chameleon invisible to predators by changing its skin color, does this on particular occasions by "selecting" (that is, producing) a particular color; and it is this secondary sense of selection which mediates the communication of the proper function to that color.[23]

---

[20] *Language, Thought, and Other Biological Categories*, p. 14; for Millikan's discussion of the following example, see *Language, Thought, and Other Biological Categories*, pp. 40-42, and *White Queen Psychology*, pp. 56-57.

[21] *Language, Thought, and Other Biological Categories*, p. 49.

[22] *White Queen Psychology*, p. 133.

The expanded definition of proper function Millikan introduces to underwrite her anti-exaptationist position reiterates and extends this notion of derived proper function in the sense that this expansion, too, depends on the idea that a proper function can be communicated to something else by a sort of secondary selection. Specifically, Millikan suggests that not just the relation of production, but any other relation between two things which has to be mentioned in the course of explaining the normal operation of one of them is sufficient to ground a proper function in the other one. A *normal* explanation explains the characteristics or performance of something not in terms of what is statistically prevalent in similar things, but rather in terms of what historically accounts for the reproduction of the characteristic or performance of similar things.

> Suppose now that in giving the Normal explanation for performance of one of the proper functions of one trait of an organism, one has to mention as a Normal condition for this performance the presence and action of some other reproduced structure or trait of the organism. Let us count the contribution of this second structure or trait to the proper performance of the first as one of the second structure's proper functions.
>
> If we count proper functions that way, the sea turtles' flippers will have digging as one among their proper functions, since they must be mentioned in explaining how the hole-digging programs in the turtles' nervous systems Normally work.[24]

---

[23] Millikan herself does not agree with this interpretation. On her view, the introduction of derived proper function is merely a terminological innovation intended to make clear the range of things to which the initial definition of direct proper function applies, and not an extension of that definition to encompass things that would otherwise not have had proper functions (Millikan, personal communication). For her most recent published discussion of derived proper function, see "A Bet with Peacocke," in Cynthia MacDonald and Graham MacDonald, eds., *Philosophy of Psychology* (Cambridge: Blackwell, 1995), pp. 285-92.

[24] *White Queen Psychology*, p. 48.

The refined definition of proper function reads as follows:

> Being "built" by natural selection is sufficient for proper function, being maintained by natural selection is independently sufficient, and having been utilized by other structures built or maintained by natural selection is also independently sufficient for proper function (*ibid.*, p. 49).

The last clause here reiterates the basic notion of derived proper function, that is, that the proper function of one thing can be grounded in a proper function, and thus a selection history, enjoyed by something else. But the relationship required for this grounding has been weakened from a production relation to a mere utilization relation. I shall use the term *expanded proper function* to designate proper functions that depend on such a utilization relationship.[8]

Millikan's introduction of the notion of derived proper function ensured that even novel things or novel uses of things could have proper functions. This is already a move in a nonpluralist direction, since it means that things that do not have a selection history of their own may still have proper functions. Similarly, the move of expanding the original definition of (direct) proper function brings most exaptations back into the fold of adaptations. Taken together, these moves amount to bringing system functions back into the fold of proper functions, and thus constitute an implicit rejection of pluralism in the theory of function. But Millikan does not appear to recognize the tension between the implicitly assimilationist move represented by derived and expanded proper functions, on the one hand, and the explicit commitment to a pluralist function theory, on the other.

The underlying cause of this tension is the equivocation on the term 'selection' first introduced by the notion of derived proper function. As we noted earlier, the primary, biological conception of selection involves a history of differential survival and reproduction, contingent upon successful performance. This history is essential, because it is what gives proper function its distinguishing feature of normativity. Furthermore, Millikan is quite explicit that this means the first instance of successful performance by something does *not* count as the exercise of a proper function. But since *derived* proper functions can be had by novel things, it appears they *are* established for the first performance, without any history of reproduction at all. Moreover, it appears that the thing produced does not have to perform successfully even once.

---

[8] A further point supporting my claim that derived and expanded proper functions are pretty much the same thing is that there is no clear dividing line between production relations and utilization relations. For example, in the case of linguistic tokens, they are, as events, produced by the speaker; but in another sense, they are utilizations of already existing natural-language devices.

> Indeed, a thing that bears no resemblance to any can opener previously on earth—suppose it has been designed in accordance with a totally new principle—may still *be* a can opener, and may be one despite the fact that it doesn't work.[38]

Here, as in the case of language tokens, the intentions of the designer are sufficient to endow the device with the proper function of opening cans, even if it never works. As we noted above, this intentional production represents a sort of secondary selection. The model for this kind of selection is clearly deliberate conscious choice, but that is not required. When the chameleon's skin-coloration mechanism "chooses" a novel coloration, that, too, is instantaneously endowed with a proper function. So when we are dealing with *direct* proper functions, 'selection' means an extended history of reproduction contingent upon successful performance; when we are dealing with *derived* proper functions, it may mean a single instance of production by something else, regardless of whether the thing produced performs successfully or not. In short, selection in the case of derived proper functions does not necessarily have *any* of the characteristic features of selection in the case of direct proper function.

This is also true of expanded proper functions. They can be established in cases where the thing utilized is novel in exactly the same sense that the chameleon's new shade of brown is novel. For example, I once tightened a loose screw on my garage door with the key for my safe-deposit box. It is quite conceivable that nobody has ever used such a key for this before or since. But a normal explanation of how the cognitive mechanism that underwrites dealings with screws performs its proper function will have to mention the utilization of screwdrivers, knives, fingernails, coins, and whatever else people sometimes use for this purpose. So my utilization is sufficient to establish an expanded proper function for the safe-deposit box key. It also seems that the utilized thing does not have to perform successfully even once. Suppose I use something bearing no resemblance to anything that has ever been used to tighten a screw before. If it were something I had designed and produced, my intention would endow it with the relevant derived proper function even if it did not work.

---

[38] *White Queen Psychology*, p. 22.

But by parity of reasoning, my intention to utilize an already existing thing ought to endow it with the relevant expanded proper function even if it does not work. So here again, the secondary selection constituted by the utilization relation does not exhibit either of the main features of selection in the primary, biological sense.[27]

This equivocation should alarm Millikan and her adherents, because it makes the conception of proper function internally inconsistent. To put the point bluntly, the introduction of derived and expanded proper functions means that proper function in general both does and does not essentially involve a selection history in the primary biological sense, and consequently it both is and is not normative. The inconsistency manifests itself in several ways. First of all, lots of things that do not intuitively seem to have proper functions at all—things Millikan herself pretty clearly does not *want* to have proper functions—will turn out to have them. For example, noses will have the proper function of supporting eyeglasses. This is because eyeglasses have a direct proper function—correcting faulty vision with lenses, let us say—and in order to explain how they normally work, you must mention, among other things, how they are supported in front of the eye. In short, eyeglasses *utilize* the nose, so noses have the expanded proper function of supporting eyeglasses. Similarly, shoulders have the expanded proper function of holding up clothes. Millikan also resists the suggestion that the mouths of dogs have the proper function of carrying things. But this case is structurally identical to that of the hind limbs of turtles, since in order to explain how the mechanisms in the dog's brain which are responsible for the carrying behavior normally work, you would have to mention its mouth.

Millikan thinks she can avoid this proliferation of unwanted proper functions by insisting that an appropriate level of abstraction must be observed when specifying them. If the initial specification is abstract enough, the troublesome cases disappear because it is no longer necessary to mention the utilized item or trait. For example, Millikan maintains that:

---

[27] This is as good a place as any to point out that in fact normal explanations in cases of utilization typically appeal to the current capacities and dispositions of things rather than just to a history of what actually has been used for certain tasks. In the case of screws, for instance, a normal explanation of how people deal with them would explain that typically a screwdriver is used, but failing that, anything which is thin, flat, rigid, and not unduly subject to breaking or splintering is used. This is further evidence for my claim that in introducing expanded proper functions, Millikan is implicitly trying to assimilate system functions to proper functions.

> ...[N]o mention of the dog's mouth need be made in giving a Normal explanation of how the general learning mechanisms in the dog that are responsible for its newspaper-retrieving behavior have historically worked.[38]

Similarly, if we simply talk about the theory of correcting faulty vision with supplementary lenses in an abstract sort of way, we need never mention any of the actual artifacts that have been created to achieve this, and consequently need never mention any of the various ways they have been supported.

But this attempt to evade the consequences of expanding the definition of proper function is clearly ad hoc. To begin with, we could also avoid mentioning the turtle's hind limbs or the chameleon's novel shade of brown by talking in a more abstract way about the general cognitive mechanisms or skin-coloration mechanisms involved. But when Millikan first introduced the notion of derived proper function, it was precisely *in order to* make functional analyses available at these very concrete and fine-grained levels of specification. Moreover, it is just not the case that all legitimate functional explanations move at the high level of abstraction Millikan recommends. Rather, there is a whole range of levels of abstraction on which such explanations may move, and which level is appropriate depends on our specific explanatory interests.

For example, we may have an interest in explaining the array of vision-correcting artifacts available, and the history of their becoming available as new technologies arose. There are salient differences among these artifacts with regard to support options: monocles are wedged into the eye socket, lorgnettes are held in the hand, pince-nez are supported by the nose with the help of a spring, eyeglasses are supported by the nose and ears, and contact lenses rest directly on the cornea. These differences concern *direct* proper functions of subparts of these artifacts. For example, the spring in a pince-nez has the direct proper function of gripping the nose, that is, it has been reproduced because some of its ancestors did that successfully. But in giving an explanation of how that spring normally works, and how it normally contributes to the direct proper function of the pince-nez as a whole, you must mention the utilized nose, which will then have the expanded proper function of supporting the pince-nez.

Similar considerations apply in the case of the dog's mouth. Millikan's argument here depends on assuming that the dog's carrying behavior is not underwritten by a special wired-in cognitive mechanism, but is generated entirely by general learning mechanisms. Since I did not find this assumption antecedently plausible, I consulted a professional dog trainer. She informed me that in the retrieving breeds and Rottweilers, puppies as young as six weeks spontaneously carry things around in their mouths, and that in these breeds especially this is not considered to be learned behavior, but rather a genetically determined behavioral disposition. Furthermore, she pointed out that in all dogs there is considerable selective pressure for carrying things in the mouth, since that is how puppies and food are transported.[29] So, if there is a special wired-in mechanism involved, either in all dogs or in some breeds, at the very least, explaining how it normally works will require mentioning the dog's mouth, just as wired-in cognitive mechanisms for digging require mentioning the turtle's hind limbs. But even if there is no such special cognitive mechanism in the case of dogs not selectively bred for retrieving, we may still want to explain how the more general mechanisms get specialized developmentally so that the dog learns how to *carry* things, as opposed to learning any of the many other things it is capable of learning; and that will still require mentioning the dog's mouth. The important general point here is that we do routinely have an explanatory interest in the cognitive structures and behavior of animals at this level of concreteness.

---

[38] *White Queen Psychology*, p. 48.

So Millikan is stuck with a distressing proliferation of proper functions if she insists on the expanded definition of proper function. This proliferation is all the more distressing because neither expanded nor derived proper functions behave at all like proper functions as described elsewhere by Millikan. The hallmark of proper function, according to Millikan and her adherents, is normativity. This means that something with a proper function is *supposed* to be able to perform it, even though it may never do so, or is incapable of doing so. Indeed, since normal does not mean statistically average, something with a proper function is still *supposed* to perform it even if the vast majority of things in its lineage do not or cannot perform it. This strong normativity is possible *only* because proper functions are established historically for lineages of things, not for single exemplars. In other words, the reproductive history of the lineage sets the standard that exemplars are supposed to achieve. If they fail to achieve it, they are malfunctioning—failing to do something they should be able to do, not just failing to do it *tout court*.

---

[29] Nanci Newton, personal communication.

Derived and expanded proper functions, by contrast, can be established for single exemplars. But on pain of further complications, the derived proper functions had better lapse with that thing or that use. This is most obvious in the case of language tokens or artifacts that get proper functions from their user's intentions. For example, if I say "turn left" and intend it to mean "turn right," then that token of 'turn left' has the derived proper function of meaning "turn right." But we surely do not want to say that from now on whenever I or anyone else produces a token of 'turn left' that it means both "turn left" (its direct proper function in any case) and "turn right." The problem here is not that the two meanings are in conflict, but that in general words do not permanently acquire new meanings by fiat of an individual speaker in this way. 'Turn left' *might* permanently acquire the meaning "turn right" as a *direct* proper function if I started routinely using it that way, and other people picked up the practice. This is, after all, how language evolves. But this process is a history of selection on the biological model, with reproduction contingent upon success. It depends on many speakers, and the intentions of any one of them have no special weight.

The case is similar for artifacts, which are as individuals subject to all sorts of idiosyncratic uses that are not then established as proper functions for other artifacts in the same lineage. For example, I have used table knives for all of the following: peeling garlic cloves (you hit the clove with the handle of the knife to loosen the skin, an operation which is perhaps even more commonly performed with the flat of a heavier cooking knife); opening cocoa tins; transplanting seedlings; removing potbound houseplants from their pots; removing crumbs from the tablecloth; and tightening or loosening screws. Now, we certainly do not want to say that in virtue of endowing one or another table knife with these expanded proper functions that I simultaneously and magically made *all* table knives into garlic peelers, screwdrivers, dibbles, and so on. But nor do we want to say that I permanently endowed any individual table knife with any of these functions. This would mean that of two indistinguishable table knives, for instance, we might have to say that one was a screwdriver but not a garlic peeler, and the other was a garlic peeler, but not a screwdriver. Again, the problem here is not that the invisibility of the individual histories of things like table knives makes it impossible to tell what derived or expanded proper functions they might have. The problem is that such cooptations of artifacts are transient reassignments of function based not on the history of the thing, but on its current capacities and dispositions. Thus, the very same table knife that functions as a screwdriver today will work just as well as a garlic peeler tomorrow, or a dibble next week. So, in the case of things less evanescent than language tokens, derived or expanded proper functions must be regarded as established only for a particular occasion of use, and not as ongoing functions of enduring individual things.

It might seem that derived or expanded proper functions in biology are established for lineages of things; for example, the hind limbs of turtles. But this is an illusion. In the first place, even in biology, these proper functions *can* be established for individual things or uses of things. For example, suppose the chameleon's novel shade of brown is never again produced by that or any other chameleon. It will still, on its unique occasion of production, have the relevant derived proper function. But even in the cases where derived or expanded proper functions apply to large groups of similar things, the instances are essentially separate. For example, if a particular turtle's cognitive mechanisms for some reason resulted in its digging holes with its forelimbs rather than its hind ones, then its forelimbs would have the expanded proper function of digging, and, unlike other turtles, its hind limbs would not. This shows that the expanded proper function is re-established for every turtle (and perhaps even on every digging occasion, since a turtle might conceivably dig sometimes with its forelimbs and sometimes with its hind ones). So if derived or expanded proper functions do apply to groups, this is a matter of aggregate statistical averages.

But if derived and expanded proper functions are essentially transient and apply to groups of things only as a matter of statistical average, this casts grave doubt on their alleged normativity. If they really are *proper* functions of some kind, then they ought to be normative, since normativity is supposed to be the hallmark of proper functions. But norms are commonly understood as standards applicable to *groups* of things. In light of the common usage of this term, it would be a strange norm indeed that was applicable only to a single thing, or, worse yet, to a thing only on a single occasion. But most derived and expanded proper functions, and perhaps all of them in the realm of artifacts and language, are applicable only to individual things or occasions of use. Furthermore, Millikan has stated repeatedly that the normativity of proper function is not reducible to statistical averages. Statistically speaking, most sperm do not fertilize an ovum, and yet that is their direct proper function—what they are supposed to do. So even in the biological cases where groups of things all have the same derived or expanded proper function as a matter of statistics, there is no norm of the sort that defines proper function in general. But then it is hard to see what legitimacy there is in calling them *proper* functions at all.

On the other hand, derived and expanded proper functions do look very much like system functions. They are not normative, and they depend on the current capacities and dispositions of things, as we saw above. System functions are also typically established for individual things or uses of things, because functionality in this case depends not on the history of the thing itself, but on the containing system. So you can create, delete, or change system functions simply by installing, removing, or changing the containing system. Because they are not normative in the first place, this transience is not problematic for system functions the way it is for proper functions.

The introductions of derived and expanded proper functions are dangerous moves for Millikan. They make the overall theory of proper function internally inconsistent by attempting to assimilate system function, a type of function which Millikan herself elsewhere acknowledges as distinct and independent. This problem could be ameliorated by scrapping the expansion of the original definition of proper function. This would constitute an acknowledgement of the independent status and significance of exaptations in Gould and Vrba's sense. But if I am right that the trouble begins already with the introduction of the notion of derived proper function, that, too, will have to be reconsidered. This would have much more serious and far-reaching consequences for Millikan's overall theory, since derived proper functions are central to her account of intentionality, among other things. But the advantage of making these revisions would be a straightforward distinction between proper-functional adaptations and system-functional exaptations. The question to ask would be simply whether the performance in question does or does not have a selection history fitting the biological model of reproduction contingent upon success. If it does, the function is a proper function (adaptation); if not, it is a system function (exaptation).

But this brings us back to Millikan's other objection to Gould and Vrba's exaptation manifesto—the objection that even though most of the examples mentioned by these authors may have begun as exaptations, these system functions have by now been transmuted into proper-functional adaptations in virtue of a subsequent selection history acting on the exapted trait. In other words, even if Millikan cannot directly reclaim system-functional exaptations as proper-functional adaptations via derived and expanded proper functions, she can reclaim most of them after the fact by showing that an exaptation is typically only a transient initial phase in the creation of an adaptation. This is not a move to assimilate exaptations to adaptations, but rather to reduce the significance of exaptations to a minimum while still allowing that they are distinct phenomena. The argument here seems to be that when an exaptation occurs it is either selected *against* because it is deleterious, in which case it disappears altogether, or it is selected *for* because it is beneficial, and rapidly becomes a proper-functional adaptation.[30] But this argument begs a crucial question. Are there cases of *ongoing* exaptations—exaptations that do not either disappear or get transmuted into adaptations by direct selection? If there are such ongoing exaptations, how important are they as a phenomenon of function?

These questions can be answered by reexamining some of the previous examples of purported exaptations in light of the simple criterion laid out above. Most of these examples concern cases where something originally built and maintained for one function takes on another function in parallel—for example, the hind limbs of turtles, the wings of (some) herons, and the mouths of dogs. Here, the crucial question is whether the trait is now being maintained in part because of the additional function. In other words, is it now being selected for in parallel with the original function, or is it just riding along? Would the additional function alone keep the trait going if the original function dropped out? In the case of the hind limbs of turtles, it seems to me the answer is clearly "yes," thus bearing out Millikan's claim that excavation is now a proper function of those limbs. This is because under current environmental conditions turtles have to bury their eggs to save them from predators. In the case of the dog's mouth, the answer would depend on how strong the selection pressures are for carrying puppies and food around. Suppose (very implausibly, of course) that dogs no longer needed their mouths for eating, would the jaw structure and dentition be maintained for carrying purposes? Here, again, I suspect the answer is "yes," especially since dogs have no other anatomical structures remotely exaptable for carrying. In the case of mantling in herons, on the other hand, the answer is almost certainly "no," thus vindicating the proponents of exaptation. While mantling is obviously a useful thing to be able to do, the fact that only one species of heron does do it indicates that the selection pressures favoring it are not very strong. Since wings of birds that do not fly rapidly become vestigial, the pressure would have to be strong to maintain wings suitable for mantling in a flightless heron. So the wings of (some) herons are an ongoing system-functional exaptation for mantling.

---

[30] For this argument, see *White Queen Psychology*, pp. 46-50.

The case of the mantling heron shows that there are indeed ongoing exaptations. But in the biological realm, they appear to be rare. In the realm of artifacts, however, ongoing exaptations of various sorts are ubiquitous. For example, people routinely stand on chairs to change light bulbs or reach things on high shelves. But chairs are not reproduced for this purpose. Moreover, there are artifacts that do have the relevant function as a proper function, namely, step stools and stepladders. So if chairs were no longer manufactured for their proper function of supporting seated humans, the most plausible consequence is that they would just disappear from the artifactual repertoire of our culture, while step stools and stepladders would proliferate. Another example I have already mentioned is the numerous exapted uses of table knives. As in the case of chairs, table knives are not reproduced for any of these uses, nor is it at all likely they would be if their proper-functional uses were to lapse. In some cases, there is already an artifact with the relevant function as a proper function—for example, screwdrivers, dibbles, garlic peelers. In other cases, there is no such artifact, but there is a plethora of artifacts that can easily be exapted. For example, cocoa tins can also be opened with the handle of a spoon or a fork, a screwdriver, or even some types of bottle opener. But either way, the existence of serviceable and readily available alternatives means that there would be no selection pressure at all for reproducing table knives for any of these purposes.

Having seen a few examples like this, the reader will undoubtedly be able to relate dozens more from her own experience. So unlike the biological realm, the artifactual realm is replete with ongoing exaptations of things. There are some cases of such system-functional exaptations being transmuted into proper-functional adaptations, but these appear to be as rare in the artifactual realm as cases of ongoing exaptation are in the biological one. An example is pipe cleaners, which were exapted for use in home crafts, especially those involving children, because they could easily be bent into the shapes of fuzzy little animals. Soon they became available in a variety of bright colors for this purpose. You can still find their much modified descendants (longer, fuzzier, and in every color imaginable) in stores which carry craft supplies (although last time I looked they were going under the name 'chenille stems' rather than 'pipe cleaners'). In short, although both the biological realm and the artifactual realm display the same two types of function, they differ markedly in the distribution of these types across the range of functional phenomena. More to the point for my purposes is that in *both* realms there exist ongoing exaptations, and in the artifactual realm this is the predominant pattern exaptations follow. So the second string of Millikan's bow is defective as well. She cannot construe system-functional exaptations as merely a transitory phase in the establishment of proper-functional adaptations, and thus as a nugatory phenomenon, in view of the prevalence of ongoing exaptations among artifacts.

The failure of these attempts to absorb or trivialize system function in relation to proper function show that it really is a distinct type of function, and that it has an important role to play in analyzing functional phenomena. In order to consolidate the second part of this claim, the next section will examine system function and proper function in artifacts in more detail. This is an important exercise not only for the present purpose of showing that a pluralist theory of function must give system function and proper function equal roles, but also for general purposes. Function theory as it has developed in the last few decades has never been applied specifically and systematically to an understanding of the functionality of artifacts. So the next section will also begin to fill this lacuna in the philosophical literature on function.[31]

---

[31] In *Darwin's Dangerous Idea: Evolution and the Meanings of Life* (New York: Simon and Schuster, 1995), Daniel Dennett also vigorously criticizes the notion of exaptation. His discussion falls into two parts. In the first, he addresses the notion of a spandrel, a precursor notion to exaptation introduced by Gould and Lewontin (*op. cit.*). This part has been dealt with thoroughly by Fred Gifford—"Why Not Spandrel Rather Than Adaptation?"—paper presented at the Eastern Division Meeting of the American Philosophical Association (Atlanta, GA; December 1996). The second part addresses the later notion of exaptation as formulated by Gould and Vrba (*op. cit.*). Dennett's arguments here are essentially the same as Millikan's, and are to be rejected on the same grounds as hers.

He begins by quoting Gould and Vrba to the effect that an exaptation is any trait that has a current function that was not selected for over the course of evolutionary history, either because the trait arose as a nonfunctional part of the organism, or because it was originally devoted to some other function. But Dennett claims this is a nugatory phenomenon.

> ...[A]ccording to orthodox Darwinism, every adaptation is one sort of exaptation or the other—this is trivial, since no function is eternal; if you go back far enough, you will find that every adaptation has developed out of predecessor structures each of which either had some other use or no use at all (p. 281)

Thereafter, Dennett repeats the slogan that all adaptations are exaptations at intervals. This move is essentially the same as Millikan's claim that exaptations are typically just transitory initial phases of adaptations. The idea, in both cases, is that exaptations are of negligible significance except as a minor part of the theory of adaptations. But like Millikan, Dennett neglects to ask whether there are any ongoing exaptations that do not get transmuted into adaptations in this way, and so enjoy an independent status. So his argument is subject to the same rebuttal as

III. PROPER FUNCTION AND SYSTEM FUNCTION AMONG ARTIFACTS

Artifacts have proper functions. These are the functions we are most likely to describe in answer to questions like 'What is that for?' or 'What is that?', and they are often reflected in our common names for things—for example, vegetable peeler, light switch, driveway, screwdriver, soap dish, clothes hanger, bookmark, and so on. Artifacts get these proper functions by a process analogous in basic respects to the natural-selection process by which biological traits get theirs.

The first step in the biological process is that a new trait arises by mutation or as the by-product of other traits. Alternatively, an exist-

---

Millikan's. But in addition, Dennett is making a rather elementary mistake here in moving from 'every adaptation developed out of an exaptation' to 'every adaptation is an exaptation'. That *X* developed out of *Y* does not ordinarily license the identification of *X* with *Y*. We do not say that because every bird developed out of an egg, every bird is an egg. The reason is that eggs are a distinct phenomenon, very much unlike adult birds in many ways, and closely related to some phenomena (for example, mammalian ova) in ways that adult birds are not. So to say that adaptations are exaptations is not warranted.

Dennett also sometimes says, however, that exaptations are adaptations (for example, p. 390), and he has independent reasons for this reverse formulation. He starts with an example: imagine a vending machine originally designed to accept United States quarters, but equally capable of accepting Panamanian quarter balboas because they are the same size, weight, and so forth as quarters. Imagine this machine is moved to Panama and used to accept quarter balboas. The question is whether its new operation in Panama is a proper-functional adaptation or not. Dennett claims that it is.

> The two-bitser was originally designed to be a detector of U.S. quarters. That was its "proper function" (Millikan 1984) and, quite literally, its *raison d'être*.... This historic fact licenses a way of speaking: we may call the thing a two-bitser, a thing whose function is to detect quarters, so that *relative to that function* we can identify both its veridical states and its errors.... On its very debut as a q-balber, after ten years of faithful service as a two-bitser, is the state it goes into when presented with a quarter-balboa a *veridical* detection of a quarter-balboa, or might there be a sort of force-of-habit error of nostalgia, a mistaken acceptance of a quarter-balboa *as* a U.S. quarter?... What must make the difference, of course, is whether it was *selected for* its capacity to detect quarter-balboas (agreeing with Millikan 1984). If it was so selected (by its new owners, in the simplest case), then even if they forget to reset the counter, its maiden move is the veridical acceptance of a quarter-balboa. "It works!" its new owners might exclaim delightedly (p. 406-07).

The point here, clearly, is that to exapt something is to select it for a particular use; and since adaptation is defined as the end result of selection, an exaptation is an adaptation, *quod erat demonstrandum*. But Dennett is trading here on the same equivocation on the term 'selection' which vitiates Millikan's derived and expanded proper functions. Specifically, in the case of adaptation, 'selection' refers to a history of reproduction contingent on some number of successful past performances; while in the case of exaptation, 'selection' means adoption for a particular use on a particular occasion, regardless of whether that use is successful or is repeated. So Dennett's proposal that exaptations are adaptations must also be rejected ing trait may be used for a new operation. Similarly with artifacts, the first step is either the production of a prototype by an inventor or designer, or someone puts an existing artifact to a new use. In biology, if the new trait or use of a trait is successful in its performance, and its success contributes to the reproductive success of its possessor, it thereby ensures its own reproduction as well. Similarly with artifacts, if the new artifact is successful it will be reproduced; initially, perhaps, for use by the inventor or designer, but later for use by other people. In the cultural milieu, this history of reproduction contingent upon success shows up as a history of manufacture and distribution by trade or sale. In the case of new uses of existing artifacts, they begin to be manufactured in whole or in part for the new market. The example from the previous section of pipe cleaners coming to be manufactured specifically for the crafts market illustrates this variation on the selection process.

There is one point here on which it will seem to many people that there is an important disanalogy between the biological and the artifactual cases, and that is the fact that a novel artifact is the product of intelligent thought and rational planning, whereas a novel biological feature is the product of purely mechanical forces. But does this constitute a decisive difference in the process of reproduction contingent upon success? I do not think so. Design rather than happenstance in the initial production may well increase the likelihood of successful performance and thus of reproduction; but this is certainly not guaranteed. Many prototype artifacts do not work at all, or do not work well enough for even their designers to pursue reproduction. Many more fall stillborn from the manufactory, as it were. They perform as their designer intended, but it turns out that this performance is not useful enough to anyone else to bother with reproduction. The United States Patent Office has on file thousands of designs and inventions that have never been manufactured or marketed for this very reason. Or sometimes the thing performs as expected, but this performance is actively undesired by society. An example of this is when the manufacture of certain types of weapons or insecticides is prohibited from the start. Successful performance thus does not just mean "performs as intended"; it means "performs in a way that is useful to its possessors," just as in biology.

But, the objection continues, with artifacts the process is intentionally directed not just in individual productions but over whole series of them. If the first prototype does not work, the designer will go back to the drawing board and, having learned form her mistakes, produce a more viable prototype. Surely, there is no analogue of this in biology? But in a sense there is. Paleobiologists have reconstructed the ancestry of many animals—for example, the modern horse—through a series of earlier forms. Although nature does not learn from its mistakes the way a human designer does, it still often produces a series of enhanced or more complex versions of a type of animal by sheerly mechanical means. So here again, although intention and rational planning may speed the process up and give better assurance of improved viability at any specific step, the difference is really one of degree, not of kind.

I shall leave this claim here for now. Clearly, the issues raised in this objection require more study and argument. They concern the relationship between intention and the nature of artifacts—a relationship which is much more complex than ordinarily thought, and is fraught with controversy to boot.[22]

There is another difference between the biological case and the artifactual case worth mentioning. With artifacts, reproduction may be contingent to some extent on *perceived* success, not necessarily *real* success. Take electric bug zappers, for example. Most people take it that their proper function is to reduce the number of mosquitos in their vicinity, but it is debatable whether they really do this. They do indeed kill thousands of flying insects, but by some accounts only a negligible number of these are mosquitos. Worse yet, by some accounts they actually attract mosquitos (not to mention other flying insects) to their vicinity, thus paradoxically reducing the population in your neighbor's yard while increasing it in yours. Despite repeated reports of these findings in the popular media, millions of these devices are manufactured and sold in the United States every year. In this case, as in many others, the perception of success is, of course, subject to manipulation by advertising.

The question is: Do we want to say that artifacts reproduced largely on the basis of such illusory success have the relevant proper functions or not? Alternatively, should we say that they do have a proper function, but just not the one they are popularly perceived as having? I think this second tack is the right one to take. Bug zappers *are* successful at killing lots of flying insects. It is that real success which underwrites their reproduction, because it allows the illusion that they are killing a lot of mosquitos, and thus ridding the vicinity of them, to take root. Similarly, the patent medicines of yesteryear often did not have any real effect on the ailments they were advertised to cure, but they did really make people feel better due to the inclusion of psychoactive substances such as caffeine, alcohol, or coco extract. In other words, the artifact case is the same as the biological case in that reproduction is contingent on a successful performance, and it is that performance which establishes the thing's real proper function. In the case of bug zappers, the real proper function is to kill lots of flying insects, and in the case of some patent medicines it is to make people feel better. But people can be, and often are, mistaken about what it is an artifact does successfully; and this leads to a mistaken attribution of proper function. The twist here is that this mistake may increase a reproduction process that is already underway—some people would buy bug zappers just because of their real success in killing flying insects indiscriminately, but many more people buy them because of the mistaken attribution to these devices of the proper function of ridding the vicinity of mosquitos, specifically. There may be biological cases that are similar to this. For example, secondary sex characteristics may or may not be linked to real reproductive effectiveness. But if enough prospective mates think they are, they will be reproduced for that reason alone.[23]

---

[22] For a more extended treatment of these issues, see Dipert.

So artifacts have proper functions that they acquire in basically the same way as biological traits acquire theirs. Let us now consider changes of proper function. There are two scenarios here. On the one hand, a change of proper function may mean that additional proper functions are layered over the original one, so that the thing ends up with multiple proper functions. On the other hand, one proper function may replace another one entirely. A biological example of acquiring multiple proper functions is the hind limbs of turtles acquiring excavation as a proper function in addition to locomotion. Acquisition of additional proper functions is very common among artifacts. Some of the most interesting cases involve what Michael Schiffer[24] calls *sociofunctions*. These are functions of a social, communicative nature, as opposed to functions of a technological nature (*technofunctions*, in Schiffer's terminology). For example, all dresses have the technofunctions of covering the body, keeping the wearer warm, and so on. But wedding dresses have acquired in addition the proper sociofunction of manifesting the wearer's social role, and designer dresses similarly manifest the wearer's socioeconomic status. These sociofunctions are usually secondary proper functions acquired long after the original function has been established. This original function may be either a technofunction or a sociofunction. Jewelry has an original sociofunction of decorating the body, but wedding rings have since acquired the additional sociofunction of communicating the wearer's marital status. The secondary function may also be a technofunction. For example, eighteenth-century living rooms often had tall mirrors set into all the walls. These were not just for reflecting the inhabitants, but to increase the ambient light level at night by reflecting back the light of candles and lamps.

---

[23] I owe this example to Cummins.

[24] *Technological Perspectives on Behavioral Change* (Tucson: Arizona UP, 1992).

The second type of change of proper function occurs when the original proper function is replaced wholesale by a new one. In biology, an example is the panda's "thumb." In this case, a bone that was originally a small proper-functional part of the wrist—the radial sesamoid bone, to be exact—became specialized for grasping through enlargement and rearranged musculature, and no longer functions in its old role. Similarly, the wings of penguins now have the function of helping them swim, not fly. This replacement of one proper function with another is common among artifacts. As in the biological case, it normally involves some modification of the artifact. For example, you can buy kits for converting a plastic soft-drink bottle into a bird feeder. The bottle is inverted, a hanger is inserted at the top, and a feeding platform with perches is attached at the bottom. The bottle then functions as the seed reservoir. So, in effect, the bottle is taken as raw material for a manufacturing process that takes advantage of some characteristics the bottle already has in virtue of the original manufacturing process. Similarly, people regularly convert old tires into planters or sandals or swings. Typically, the remanufactured item loses its ability to perform its original proper function, sometimes as a result of the modification (the soft-drink bottle), but more usually as the result of wearing out (the tire). The remanufacturing process then results in the acquisition of a new proper function, sometimes for the thing as a whole (the tire planter), and sometimes as a proper-functional part of some other device (the soft-drink-bottle bird feeder).

But there is one problem here, which arises out of a disanalogy with the biological case. When artifacts are remanufactured into different artifacts, it is not usually every artifact of that type which is remanufactured. Not all soft-drink bottles are made into bird feeders, nor are all of them recycled into carpets or lawn furniture. But because of the constraints of genetic inheritance in biology, if one panda's radial sesamoid wristbone is a thumb, all of them will be. From this point of view, it might seem that it is only in biology that it is appropriate to speak of replacement of function. In the case of artifacts, we might do better to treat cases like the soft-drink bottle bird feeder as the addition of a proper function. Here is why. Millikan's theory of proper function dictates that although the soft-drink bottle is now completely unable to perform its original proper function, it still *has* that function because that is still what accounts for its reproduction. Similarly, a liver so badly damaged by toxic substances that it is incapable of performing its proper functions still has those functions, and *should* perform them. In short, it seems that the defining normativity of proper functions really means no proper function can ever be lost or replaced by a different one unless there is wholesale reproduction only for that new function. On this interpretation, remanufacturing of some exemplars of a type of artifact would not count as reproduction, but only as a modification of an item reproduced for another reason.

But this does not seem quite right, either. When we consider the case of the soft-drink bottles recycled into carpeting, we do not want to say that the latter retains the proper functions of the former. But it seems to be only a matter of degree of modification which separates recycling of this sort from other, less drastic, modifications. The situation is further complicated by some cases among artifacts where a version of the remanufactured item is manufactured from scratch and takes over the market. For example, the first slide guitarists used the neck of a glass bottle, carefully cut and ground down. But nowadays you can buy specially made glass cylinders in a range of sizes and styles for this purpose. Here, there is first a remanufactured type of artifact, which then gives rise to a new type of artifact altogether. In cases like this, there is an important historical relationship between the two groups of artifacts, and because of this they seem in some ways to be the closest analogue to biological cases like the panda's thumb, where there is an evolutionary history to be traced between wristbones and dedicated thumbs. But there are obvious disanalogies here as well—the panda's thumb is still clearly a reworked wristbone, whereas specially made slides are no longer reworked bottlenecks, even though they bear an important historical relationship to reworked bottlenecks and are sometimes still referred to as bottleneck slides. (This is to distinguish glass slides from similar cylindrical metal slides, which originated in the practice of using a knife blade on the strings instead of a bottleneck.) We do need some theoretical apparatus for talking about replacement of proper function in any case, in order to deal with biological cases like the panda's thumb. But it is unlikely to fit the artifactual cases where, intuitively, there seems to be a similar replacement. In cases like the bird feeder and the bottleneck slide, we face the theoretical question: How much modification is necessary, and how uniform must that modification be over the type of thing in question, before we should say one proper function has replaced another? More simply, what do we mean by reproduction in the case of artifacts?

These are actually very difficult questions, because they concern the relationship between form and function, which, like the relationship between intention and function, is complex and fraught with controversy. But a full-fledged theory of function for artifacts will have to answer such questions, since the remanufacture of artifacts to fit them for a new function is both exceedingly common and rife with disanalogies to the closest biological analogues. This may mean that some fairly serious revisions of the theory of proper function will have to be undertaken to make it applicable to artifacts as well as to the biological traits it was designed to fit.

There are also a couple of interesting special cases of change of proper function which are relatively uncommon even among artifacts, and may be missing altogether in biology. First, there are some artifacts that are designed to have proper functions serially. For example, the Vermont Country Store sells maple syrup in containers shaped like log cabins, which are advertised as converting into a child's bank once the syrup is gone. Second, changes of proper function need not involve any modification of the thing at all. Among artifacts, this happens when something is reproduced without modification entirely for a new use. The best examples of this in our culture are authentic reproductions of items from exotic or past cultures for decorative purposes or as souvenirs. Although I think there could in principle be biological analogues of these types of function change, I am at a loss for examples. In any case, these special cases must also be accommodated by any viable account of change of function in artifacts.

We can now go on to address system function in artifacts. The first point that must be made, however, brings us right back to the process of acquiring proper functions just discussed. Regardless of whether a thing is acquiring its first proper function, acquiring additional ones, or acquiring a new proper function in place of its original one, the acquisition process always starts with the institution of a system function. This is tacitly acknowledged by Daniel Dennett when he says that every adaptation starts as an exaptation; and it is tacitly acknowledged by Millikan when she says that the first successful performance in a selection history is not yet an exercise of the proper function to come. Both of them interpret this as showing the dependent and relatively insignificant status of system-functional exaptations in comparison with proper-functional adaptations. I have already argued that this interpretation ignores the existence of ongoing exaptations, especially in the artifactual realm. But another objection arises out of the prevalence of addition and replacement of function among artifacts. These phenomena manifest the essentially labile nature of functions. Functions are constantly coming into and going out of existence. This lability is simply more visible in the artifactual realm, where the rate of change is a lot faster than in biology. The important point here is that system function is the linchpin on which acquisition and change of proper function depend. If it were not for the initial institution of a system-functional exaptation, no proper-functional adaptation could ever be acquired or changed, because there would be no initial successful performance for selection to act upon. To put the point more generally, although natural selection is often referred to as an account of change, it is in fact only part of such an account. The other notion necessary to any full-fledged account of change in either the natural or the cultural realm is system function, which provides an analysis of the crucial initial phase of any process of change in terms of the subsidiary notions of containing system and system components.

From this point of view, it would be more appropriate to describe every proper function as merely an ossified form of a system function, thus reversing the relations of dependence and relative importance as assessed by Millikan and Dennett. A truly pluralist theory of function, however, would do better to regard these two types of function as complementary and of equal importance. Just as every proper function starts as a system function, so many system functions start as exaptations of something which already has one or more proper functions, and which would not be available for exaptation but for those proper functions.

In addition to the essential role of system functions in the establishment of all proper functions, there are system functions that never do get consolidated as proper functions. As I noted in the last section, these ongoing exaptations are infrequent in biology, but common among artifacts. They usually are cases where an artifact is used for something unrelated to its proper function, but is not modified for that use. So the artifact can in principle be returned to its proper use at any time, and in practice it often is. These ongoing exaptations may be classified into standardized and idiosyncratic.

Standardized ongoing exaptations are exaptations that occur as a matter of widespread cultural practice. For example, in our culture, people regularly stand on chairs to reach things. Similarly, they use screwdrivers to open paint cans, spoon handles to open cocoa tins, diluted dish detergent to find leaks in tires and other pressurized gas systems, antique flatirons as doorstops, and so on. Despite being the subject of a regular practice in this way, the artifacts in question are not reproduced for this use, so they never acquire it as a proper function. Instead, the system function is an ongoing additional function layered over the original proper function.

Sometimes the exaptation is a fairly permanent reassignment of the artifact to its system-functional role, and return to its proper-functional use is not expected in fact, although it usually remains possible in principle. For example, when you exapt a flatiron for a doorstop, you do not normally have any intention of using it as an iron again, especially since the flatiron has been supplanted in general use by the much safer and more efficient electric steam iron. But you *could* always use the flatiron again—for instance, if you wanted to demonstrate just how much safer and more efficient electric steam irons really are. In cases like the detergent leak detector, the particular quantity of detergent used in the operation is used up, and so is not available for any further use. But the rest of the detergent in the bottle can be used for its proper function.* In most cases, however, the artifact is immediately available for its proper function again once the system-functional use has run its course. You can stand on a chair one minute, and sit on it the next. Thus, these exaptations constitute a transient but repeated use of the artifact.

There are also *idiosyncratic ongoing exaptations*. They do not correspond to any recognized cultural practice, but rather represent creative uses of artifacts by individuals. Here are some examples from my own household. An antique cast-iron doorstop is a bookend; the cord of the kitchen radio is an antenna in virtue of being tied up in a loop to the handle of an overhanging cabinet; twist-ties are cat toys; an old wooden firkin and a ten-pound Planter's peanuts tin are wastebaskets; and in the vegetable garden, some tomatoes are tied to their supports with old shoestrings, and a four-foot length of foundation drain pipe helps ventilate the compost pile. Of course, it sometimes happens that several individuals separately come up with the same exaptation. For example, some time after discovering that my cat loved to play with twist-ties, I began hearing stories of other cats whose owners had made the same discovery. It also happens that an idiosyncratic exaptation becomes a standardized one by a process of diffusion. Indeed, this is how all standardized exaptations must start. For example, an acquaintance of mine reported giving twist-ties to his cats after hearing from his mother that the family cat liked to play with them. This diffusion process is often helped along by reports in the popular media. For example, *Organic Gardening* magazine has a regular column called "Garden to Garden" devoted to reports from readers about innovative techniques and solutions they have found, and these frequently involve novel exaptations.

---

* Artifacts that get used up or are otherwise ephemeral—soap, food, perfume, ink, electricity, dry ice, and so on—are a distinct type of artifact which is all too rarely discussed. Such artifacts do have some special features that would have to be accommodated by any theory dealing with the general ontology of artifacts, but as far as function goes, they have the same range of proper and system functions as any other type of artifact.

In the last section, I noted that there are some ongoing exaptations in biology—for example, the mantling use of the heron's wings. These correspond to standardized exaptations of artifacts, the only difference being that they are species-wide rather than merely commonplace in a local culture. These biological standardized exaptations, too, must have started as idiosyncratic exaptations by individual animals. But this analogy with biology reminds us that exaptations may also involve pressing into service some feature that hitherto has had no function at all, proper or otherwise. For example, Gould and Vrba mention the skull sutures of mammals, which help them pass safely through the birth canal (*op. cit.*, pp. 5-6). But infant reptiles and birds, who have no need of this help, also have these sutures. It appears that the sutures are a by-product of general developmental constraints, and thus originally had no function at all. But at some point they were exapted by mammals for their current use, and may by now count as a full-blown proper-functional adaptation. If we did not already have them, we would have had to invent them, as it were. In any case, the point here is that exaptations can operate on nonfunctional items. There is an important artifactual analogue of this phenomenon. It occurs most obviously when something that has no cultural function at all, like a stone or a stick, is pressed into service for a human purpose. For example, you might hammer in a tomato stake with a handy stone, or carry a serendipitous find of wild strawberries home in a large leaf. Another form of this occurs when some nonfunctional feature of an artifact is similarly pressed into service. A well-known example of this is the architectural spandrel, made famous by Gould and Richard Lewontin (*op. cit.*). This is a feature which is a by-product of other architectural decisions, and has no function of its own. But once it is there, it is routinely exapted for decorative purposes. A more commonplace example of the same phenomenon is the use of the page edges of books for decorative gilding or marbling.

The first form of exaptation of the nonfunctional is particularly important for any theory of artifacts, because it occurs at the interface between the natural and the cultural, and thus helps articulate the relationship between these realms. In terms of the sheer possibilities of exaptation, there is no significant difference between exapting a naturally occurring thing and exapting an artifact. In both cases, the primary concern is whether the thing will serve the system function for which it is needed. This is a matter of the thing's current capacities and dispositions, not of its provenance. So from the point of view of a projected exaptation, the realm of artifacts is a seamless extension of the realm of naturally occurring things. This means that system function will be a crucial notion for understanding the use of tools by nonhuman animals, who only rarely manufacture them. It will also be crucial in understanding the history of hominid tool use, which developed from the simple exaptive use of naturally occurring objects as hammers, digging sticks, and so on, to the pervasively artifactual environment we Western industrialized humans inhabit today.

This preliminary exploration of artifact function clearly confirms the need for a pluralist theory of function in order to account for the salient phenomenon of exaptation, particularly ongoing exaptation, in this realm. The generally labile nature of function which leaps to the eye in culture in a way it does not in biology shows that the correlative notions of function constituting this pluralist theory ought to be regarded as of equal theoretical importance. But a further point is now in order. The pluralist theory outlined in sections I and II, which was originally standardized on the biological realm, has shown itself to be adequate only to a very basic account of artifactual function. The artifactual realm has some special features affecting functional phenomena which appear to have at best loose analogies to functional phenomena in biology. So even a pluralist theory of function will require extension and revision if it is to accommodate artifact function as well as biological function. I suspect some of the required adjustments may prove to be far from minor.

IV. CONCLUSION

Artifacts have been woefully neglected as a topic of philosophical investigation; but this neglect is not benign. In particular, it is not benign in function theory, where a failure to give sustained attention to artifactual function has obscured the need for a pluralist, general theory of function. It has also obscured the nature and extent of the differences between artifactual function and biological function, which remain to be examined more thoroughly in future work. When finally and fully articulated, these differences may require further revision and extension of the pluralist theory defended here.

BETH PRESTON

University of Georgia

COMMENTS AND CRITICISM

MULTIPLE MEANINGS AND THE STABILITY OF CONTENT

Eric Lormand[1] provides meaning holists with a bold new response to what many people, including Lormand, take to be the main objection to meaning holism: that meaning holism leads to meaning instability. His proposal is particularly interesting because he claims not only to meet the challenge that meaning stability poses, but to transform this challenge into an argument in favor of meaning holism. "My positive aim is to show how meaning holism, even at its most extreme, can accommodate and also increase meaning stability" (51). The innovation responsible for this feat is, according to Lormand, the idea that concepts have multiple meanings. If he is right about this, and he actually succeeds in making a virtue out of what has been taken to be the main difficulty for meaning holism, then a good deal of current work in the philosophy of mind and language is deeply misguided. We shall argue, however, that Lormand does not in fact succeed because his account is inadequate in dealing with the threat that meaning stability poses for meaning holism. In the end, the intriguing hypothesis that symbols have multiple meanings has little effect on the issue of whether meanings are stable.

Meaning holism, like many -isms, is a family-resemblance notion; it covers a variety of related views about the nature of representation. A core idea in this family is that the content of a given type of symbol in a system of representation depends upon its relation to most other types of symbols in the system. This means, for example, that the English word 'bird' has the content it does not just because of its relation to some small number of prescribed words (say, 'animal', 'flies', 'wings') but also because of its relation to most other words of English ('pencil', 'tunnels', 'parka', 'galaxy', and so on). In spite of the counterintuitiveness of this result, meaning holism has many adherents in philosophy and cognitive science. Indeed, in contemporary philosophy of language, it has become something of a characteristic doctrine. The reason for this, as Lormand points out, is that it has proved exceedingly difficult to draw a line between those links among symbols which contribute to their meaning and those which do not. Assuming that some links among symbols contribute to their mean-

---

[1] "How to Be a Meaning Holist," this JOURNAL, XCIII, 2 (February 1996): 51-73.

0022-362X/98/9505/255–63

ing, the only principled view seems to be that *all* of them do; hence the pressure toward meaning holism.

Meaning holism is connected in a fairly direct way to worries about the stability of content. By the ***stability of content***, we mean the fact that two symbols can express the same content even though they fail to bear the same relations to other symbols; or, to put it differently, that the content of a symbol remains the same in spite of changes in its relations to other symbols.[2]

The point is easiest to see in reference to mental content. Suppose for the moment that thinking takes place in an internal system of representation, where beliefs and other propositional attitudes are understood as relations to structured mental representations. On this view, thoughts are much like sentences in that they are composed of simpler elements that themselves have semantic content. Call these constituents *concepts*.[3] Now, when meaning holism is applied to mental content, this means that the content of a concept depends upon its relation to most other concepts in the internal system of representation. Usually, the relations that are thought to matter are a concept's inferential relations, that is, the causal relations it bears to other representations in cognitive and perceptual processes.[4] But then, to the extent that people have different beliefs or different inferential tendencies, the inferential relations in which their concepts participate will diverge. As Lormand points out, these divergences make it extremely difficult to characterize interpersonal—or even intrapersonal—agreement and disagreement. The situation is especially graphic on the strong version of meaning holism, which Lormand adopts for expository purposes. On this version, *any* change in belief involving a given concept changes the meaning of that concept, and since the conceptual system is taken to be thoroughly interconnected, any change in the meaning of one concept changes the meaning of every other concept (55). The result would seem to be that no two people can agree or disagree with one another. Similarly, since any individual's beliefs are constantly changing, her concepts must be changing. So, it looks like, a single person could not even agree with her former or future selves. Taking this reasoning to its limit, Lormand points out that a single person could not even execute a simple deduction without equivocation, for in accepting the premises she would thereby change the meaning of the concepts involved (56).

---

[2] Notice that the first formulation concerns two token symbols and whether they are of the same content type, whereas the second concerns a symbol type (where the type is determined by the content) and conditions on instantiating that type. We shall generally let context make clear whether types or tokens are at issue.

[3] Following Lormand, we shall use bracket notation to refer to mental representations, where the content of the representation is given by the bracketed expression. For example, [bird] is the concept bird, and [some birds fly] is a representation with the content that some birds fly.

[4] See, for example, Ned Block, "An Advertisement for a Semantics for Psychology," in Peter A. French, Theodore E. Uehling, and Howard K. Wettstein, eds., *Midwest Studies in Philosophy, Volume X: Studies in the Philosophy of Mind* (Minneapolis: Minnesota UP, 1986), pp. 615-78.

These are certainly bracing consequences, and meaning holism would clearly be a far more palatable view if only these consequences could be avoided. In this context, Lormand's defense of meaning holism is especially welcome. Rather than ignore the problem of the stability of content (like most meaning holists), he argues that, contrary to appearances, meaning holism is actually perfectly compatible with the stability of content. On his account, the innovation that is responsible for the compatibility is the idea that a token symbol has multiple meanings.

Lormand introduces his characterization of multiple meanings by asking us to imagine all of the potential causal/inferential links in which a given concept might participate. In general, these are to be divided into (possibly overlapping) classes, each of which Lormand calls a *unit*. He suggests that, to a first approximation, we think of a unit "as a separable rough test for the acceptable use of that representation" (57). In Lormand's example, one unit associated with [bird] might be [feathered flying animal], another [thing called 'a bird' by mommy] and a third [thing that is similar enough to B1, B2,...,B*n*] (where B1, B2,...,B*n* are things taken to be birds). Lormand's proposal about multiple meanings amounts to the claim that each unit associated with a representation constitutes one of its meanings and that each representation may have as many meanings as it has units. We are supposed to think of *the meaning* of a concept as the collection of all of its individual meanings.

It is this innovation that leads Lormand to claim that meaning holism allows for meaning stability. "Two representations may have different total sets of units but still share meaning, if the sets overlap, or share at least one unit. Even after a change in one unit, a representation still has many of the same meanings it did before" (59). The suggestion, in other words, is to allow that the meaning of every concept changes with any change in belief, but to insist that this requires only that at least one meaning of each concept changes.

Lormand illustrates this proposal with an example involving two children who have rather different cognitive lives. Larry, we are told, has a mental representation [bird],[5] which is associated with the units [feathered flying animal], [thing called 'a bird' by mommy], and [thing that is similar enough to B1, B2,...,B*n*]. His brother, Moe, who perhaps is blind, also has a mental representation [bird] which is associated with [thing called 'a bird' by mommy] but not with [feathered flying animal] and not with [thing that is similar enough to B1, B2,...,B*n*]. Lormand's suggestion is that, despite their differences, the two children are capable of genuine agreement and disagreement precisely because they associate at least one of the same units with [bird]—namely, [thing called 'a bird' by mommy]. This really is the crucial point for Lormand: shared meanings, or units, are the basis for a certain amount of stability.[6]

It is important to see, however, that the proposal does not really establish stability in *the* meaning of concept. At best, it establishes stability in some of the meaning*s* of a concept. Moreover, there is no simple relation between two concepts sharing meanings and their having the same meaning. It might turn out, for example, that two people share only one meaning for [bird] (as in the case above), but that they share several meanings for concepts that, intuitively, are not the same. It might happen, for example, that Larry erroneously believes that birds sometimes make croaking sounds, and that Susie is afraid of them, while Moe correctly believes that Susie is afraid of frogs and also believes that frogs sometimes make croaking sounds. In this case, Larry's [bird] shares more meanings with Moe's [frog] than it does with Moe's [bird].[7] Having "the same meaning," then, is not a simple function of sharing meaning units.[8] Nonetheless, Lormand might be prepared to live without a richer notion of sameness of meaning. He might claim that he only needs meaning units to be stable, since, with the stability of these in place, he can make sense of agreement and disagreement across individuals and over time. Larry and Moe agree that Tweetie is a thing called 'a bird' by mommy, and they disagree about whether Tweetie is a feathered flying animal. Allegedly, standard versions of meaning holism cannot even provide this, so Lormand's proposal may look like a big step forward.

---

[5] Here, and below, we speak loosely of his "mental representation [bird]." Strictly speaking, we should use some more neutral locution, since the issue of meaning stability just is whether or not Larry and Moe both possess the same mental representation: [bird]. Perhaps the best strategy would be to construe claims about their "[bird] mental representations" as references to mental representations that are individuated nonsemantically.

[6] The nature of the agreement in such a case is rendered more explicit following Lormand's related suggestion that beliefs involving a concept be understood as determining sets of more specific beliefs, one for each unit associated with the concept. So if Larry thinks that Tweetie is a bird, this really amounts to his having a set of beliefs which includes the belief that Tweetie is a feathered flying animal, the belief that Tweetie is a thing called 'a bird' by mommy, and the belief that Tweetie is similar enough to B_1, B_2,...,B_n. Since Moe also associates [thing called 'a bird' by mommy] with [bird], if he thinks that Tweetie is a bird, he too believes that Tweetie is a thing called 'a bird' by mommy. Hence their agreement about Tweetie consists in their having identical beliefs that Tweetie is a thing called 'a bird' by mommy.

[7] It is particularly important to remember here that we are using the symbolism for referring to concepts in a loose way, as noted earlier in footnote 5. In particular, one cannot assume on the theory under consideration that Moe's [frog] and Larry's [bird] are less similar in meaning than Moe's [bird] and Larry's [bird].

What is more, Lormand does not stop at the claim that multiple meanings secure some amount of stability. He adds that the multiple-meaning view secures more stability than purely referential theories of content. His argument has two stages. First, he claims that so long as "a representation has a long list of meanings anyway (that is, those yielded by its units), we can simply add referents to this list of meanings" (59). "Such a theory," he says, "would at least provide all the meaning stability of a referential theory, since every meaning posited by the referential theory would also be posited by the holistic [theory]..." (59). Second, he adds that the holistic theory that incorporates referents is superior to a purely referential theory because the holistic theory "provide[s] all the psychologically-relevant meanings that purely referential theories omit" (59-60). For Lormand, psychologically-relevant meanings are more fine-grained than referents. This is because psychology is supposed to require a taxonomy of mental representations which is responsive to interesting psychological generalizations, yet coreferential representations often have different psychological consequences. Lormand's proposal is to treat the meaning units (for example [is called 'a bird' by Mommy]) as providing fine-grained meanings for concepts. According to Lormand, the resulting holistic theory is superior to a purely referential theory because shared meaning units establish points of stability over and above shared referential meanings.

In Lormand's hands, then, holism looks promising. The multiple-meanings version of holism appears to allow for a considerable amount of content stability—indeed, more stability than its nonholistic competitors. But appearances are deceiving. Lormand's proposal does not really give us what it promises. What it does, in effect, is trade the issue of the stability of the meaning of a concept for the issue of the stability of the units which constitute the various meanings of the concept. The problem with this strategy, however, is that, given a strong form of holism, *the contents of the units for a given concept are no more stable than the contents of the concepts themselves.*

---

[8] These issues are complicated considerably if, as Lormand suggests, we allow "weak" or "nondiagnostic" tests (for example, [birds are smaller than the Milky Way]) to count as individual meanings (58).

Consider the following simple case. Suppose that Larry acquires a single new belief that changes the meaning of one of the units in his concept [bird]. Perhaps he acquires a new belief about feathers, which changes the meaning of his concept [feather], and so changes the meaning of the unit [feathered flying animal] in his concept [bird]. Can this change be isolated? It looks like it cannot. Since one of the units in [bird] has changed, the meaning of [bird] changes, so, if Larry also believes that birds are animals, then the meaning of [animal] changes as well. Since [animal] changes, [human] changes. Since [human] changes, [mommy] changes. Since [mommy] changes, [thing called 'a bird' by mommy] changes. If Larry believes that [mommy's hair is similar to mine], or has some such belief, then changing the meaning of [mommy] also changes the meaning of [similar] and thereby changes the meaning of [things similar enough to B1, B2,...,B*n*]. (See figure 1.) As this simple chain of changes illustrates, the holist cannot insulate changes in meaning. Since the units themselves are representations composed of other concepts, these "unit concepts" will themselves be linked to concepts that are more immediately affected by the initial change, and thus they, too, will change meanings.

The upshot is that, on a thoroughgoing holism, changing the meaning of even a single unit will involve a change in the meaning of just about every other concept in the representational system, including all of the other concepts that are labeled as "units."[•] The multiple-meanings hypothesis has done nothing at all to diminish the threat of meaning instability. Before, when concepts were understood to have just a single meaning each, we faced the problem that a single change in belief would result in every concept changing its content. Now, with the multiple-meaning hypothesis, the problem is that with a single change in belief, every concept changes *all of its*

[•] The specific beliefs used in this example are purely illustrative. All that is needed is that all concepts are linked to each other somehow or other. The same point could perhaps be made more simply by noting that any given concept is likely to show up in the tests for nearly every other concept. As Lormand says: "A complete specification of the tests for a typical conceptual representation is likely to include virtually all of one's other representations" (58). So changing the meaning of a single concept will fairly immediately change the meanings of all the concepts to which it is connected. Since these concepts include those used in the units of a representational system, the change will lead to changes in the meanings of all the units associated with all of the concepts of the representational system.

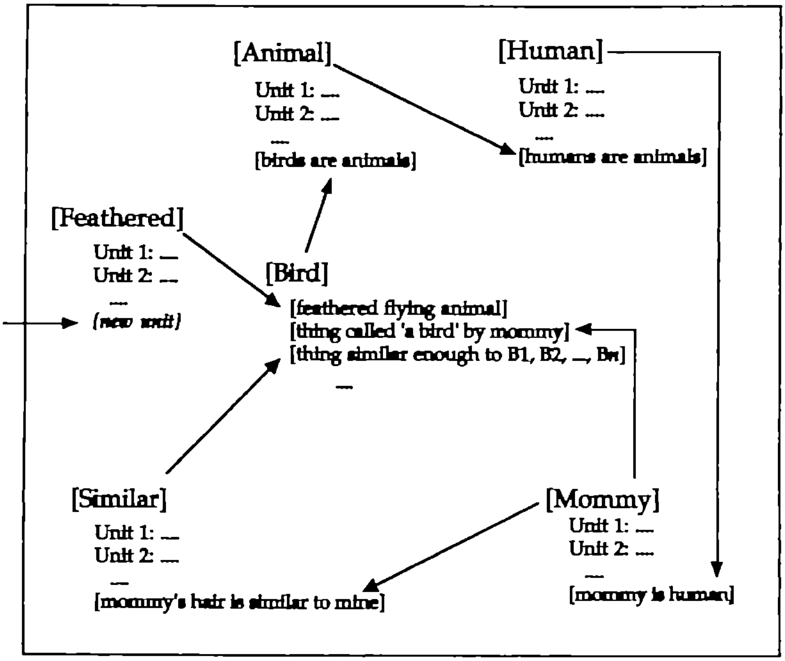


Figure 1. An illustration of how a change in the meaning of a single unit of a typical concept—[bird]—leads to changes in the meanings of all the units of that concept. Arrows indicate linked changes in meaning.

meanings. Either way, instability appears to follow from the commitment to holism.

As far as we can see, the only way that a multiple-meaning approach could contribute to stability would be if the semantic theory had two parts, one of which was not holistic. Lormand makes a suggestion apparently along these lines, in a footnote. He remarks that

> ...the meanings of all the representations in a system can be specified completely in terms of a relatively small set of semantically atomic (or simple, primitive, basic) representations in the system. A representation can have *a* primitive meaning—and so be a semantic atom while also having *other* nonprimitive meanings—and so satisfy inferential-role semantics and meaning holism. Units consisting entirely of atomic representations, then, can yield meanings which are independent of other representations in the system (58).

Insofar as we understand this suggestion, we do not think it helps.

Note that in adopting a two-component theory, Lormand abandons his advertised project of defending the strong form of holism against the objection that it leads to meaning instability. What has become of his "positive aim" of showing "how holism, even at its most extreme, can accommodate and also increase meaning stability"? But, moreover, all of the stability that is achieved is due to the nonholistic component of the theory. If this is not obvious, consider that, apart from those few concepts in the "small set of semantically atomic (or simple, primitive, basic) representations," we need to assume, in accordance with meaning holism, that the concepts in the representational system are thoroughly interconnected. Then, as before, any change in belief leads to changes in the meanings of all of these concepts, and consequently to the meanings of any unit in which one or more of these concepts figure. The only unit meanings that remain stable are those which do not contain any concept whose meaning depends upon the network. So meaning stability is still a problem exactly to the extent that the meanings of a representational system are determined holistically.

Much the same point applies to Lormand's suggestion that we might "simply add referents" to the list of meanings a concept has, thereby gaining all of the stability that a nonholist reference-based theory has. Although this move would lend some stability to the conceptual system, the stability achieved is in direct proportion to the extent to which the account is nonholistic. In any case, standard two-factor versions of inferential-role semantics already have this much stability. The interesting question is whether Lormand's proposal provides the meaning holist with a new and stronger response to the threat of meaning instability. Lormand suggests that his theory provides more stability than either a nonholistic reference-based theory or a standard two-factor theory because, in addition to the stability of the referential component, his account includes more fine-grained stability from the various stable meanings that concepts have. As we have shown, however, Lormand's theory fails to deliver these additional points of stability, since, to the extent that the account is holistic, not only will all concepts change meanings with any change in belief, but all units will change meaning as well.

In sum, Lormand does not show how meaning holism can meet the objection that it suffers from meaning instability. Multiple meanings do not secure stability, since, assuming meaning holism, the multiple meanings of a concept are themselves unstable. So Lormand's original bold claim that even a thoroughgoing holism is compatible with meaning stability—and that stability is actually a point in favor of meaning holism—is simply false. Moreover, isolating some units whose meanings are determined nonholistically does not help, since the conceptual system remains unstable exactly to the extent that it is holistic. Meaning holists need to look elsewhere for ways to meet the threat of meaning instability; Lormand's multiple-meaning hypothesis leaves the threat exactly as it was.*

ERIC MARGOLIS

Rice University

STEPHEN LAURENCE

London School of Economics

---

*The order of the authors' names is arbitrary.

BOOK REVIEWS

The Life of The Cosmos. LEE SMOLIN. New York: Oxford University Press, 1997. 358 p. Cloth $30.00.*

How many universes are there and are they actually living things? Do extremely improbable phenomena require explanation (and is this where God comes in)? Where do we come in? What about the universe? Why is it tuned to allow for our existence? Is that really improbable, or is it in fact typical? What then of life: typical or improbable, requiring explanation or not? Is Friedrich Nietzsche finally dead? What do Big Sur and Verona have that you will not find in your local shopping center? (Hint: think of Pablo Picasso or Martha Graham.) Is atomism finished and is relationism all-the-way-down a coherent alternative? Why do bright students hate physics? What do C. S. Peirce, G. W. Leibniz, Immanuel Kant, Ernst Mach, Albert Einstein, Niels Bohr, and Charles Darwin have in common? Why are contemporary philosophers so polite, anyway? Has science really shown that hidden valuables are nonlocal (if so, should Sigmund Freud not be on the list of notables)?

Clearly, the last is a typo ('hidden variables', of course; 334): the best slip, we think, of a rather large lot. We feature it to demonstrate that not all of us are that polite. It is not out of politeness that we recommend Lee Smolin's book, which treats all these questions (and even answers some); for it is brilliant and enchanting—and annoying. It is a book for Smolin's scientific colleagues as well as for the educated lay person. Most of all, it is a book for philosophers.

Smolin is an outstanding theoretical physicist, accomplished in both quantum theory and general relativity, and passionate about putting these together in a scientific cosmology that addresses fundamental questions about our universe as a whole and its inhabitants. Smolin does not think that fundamental physics, as usually conceived (the "atomism" above), can do the job. For the standard model of elementary particle interactions (and even its possible successor in superstring theory) contains too many unexplained parameters (in the case of the standard model, twenty or so, including coupling constants and the masses of the elementary particles). Instead, Smolin thinks that both quantum theory (nonlocality) and relativity (Mach's

---

* Thanks to Sherri Roush, Tom Ryckman, Lee Smolin, and Sandy Zabell for helpful discussions.

0022-362X/98/9505/264–68

principle) point to the relationism of Leibniz, which tutors us to think of the universe as an interconnected whole that can be understood only from the inside. We need to give up the chimera of a God's-eye point of view, the Newtonian (and Einsteinian) ideal of a detached, outside observer. Smolin couples his relationism with a Peircean, evolutionary perspective on natural laws, a generalized Darwinism that Smolin calls *cosmological natural selection*.

Smolin is a many-universe theorist or many universer. He conjectures that a new universe is born whenever a star collapses to a black hole. (We can guess that in every galaxy of our universe alone, there may be 10^8 black holes). Adapting the physicist John Wheeler's suggestion that certain fundamental constants of nature might be reprocessed in gravitational collapse, Smolin proposes that the free parameters of the standard model vary in the offspring universe to which a black hole gives birth. His idea is that the form of the laws remains the same but that there are small changes to the parameters in the reprocessing, in a manner analogous to the small mutations of genetic material common in biological evolution. (Despite Smolin's professions of a thorough relationism, an insiders-only view, this many-universe picture is the view of no one in particular, that is, a God's-eye view.) The values of the parameters determine a universe's ability to form stars and black holes and hence progeny, so that, over time, a typical universe plucked from the hypercosmic grab bag will be optimally tuned for the production of stars and black holes. Since our own universe appears to be so-tuned, this has at least some chance of explaining why our universe is as it is. We appear to get some understanding of life thrown in for the bargain, since the same tuning of the parameters that optimizes star formation is, Smolin claims, also highly conducive to stable atoms, organic molecules, and the tendency toward local thermodynamic disequilibrium, conditions necessary for the emergence of life itself.

What is life? For Smolin, it is a self-organized, nonequilibrium system governed by a symbolic program (like the genetic code) which can reproduce both itself and its program (156). Nonequilibrium means that energy flows through the system. According to the ideas of "critical self-organized systems," this process rapidly and inevitably creates structures (self-organization) over various scales. Thus, Smolin conjectures that the structure and hierarchy we find in the universe is the result of criticality. In his boldest move, Smolin even suggests that our universe itself is alive, a living system that made itself and is constantly on the remake. In this picture, there is no "origin" of life, nor any need for an outside maker, since life has always been. Nor, in Smolin's scenario, need we succumb to Nietzschean pessimism or nihilism, for perpetual life implies perpetual disequilibrium; so we are not doomed to a bleak, cold "heat death."

Smolin's story is not supposed to be just another origin myth (the earth carried on the back of a turtle, darkness on the face of the deep punctuated by the sudden creation of light and life). Although he emphasizes that his story is speculative, he regards it as speculative science because he regards it as an explanatory account that can generate falsifiable predictions (unlike the weak anthropic principle, which he criticizes for not being falsifiable). Cosmological natural selection not only explains the fine tuning of the twenty free parameters necessary in the "standard model" of particle physics, it also predicts that any changes (increases or decreases) in these parameters would decrease the rate at which black holes are produced. Several of these changes (for instance, changes in the masses of proton, neutron, electron, and neutrino, or in the strength of the electromagnetic field) lead to unstable nuclei and hence to a world without carbon and the carbon-based organic molecules that run the thermodynamic processes underwriting the production of black holes. In the text and an appendix, Smolin surveys eight changes in those parameters, the ones on which we can currently get a handle, and argues that each of these would lead to a decrease in black-hole production, in accordance with the theory. He also looks at modified versions of the framework adapted to superstring theory.

Both of the notions on which Smolin relies (explanation and falsifiability) are tricky and, moreover, it is by no means clear that they patrol any boundary between science and mythology (as we can see from recent, falsifiable upgrades to creationism). What, after all, does Smolin's selectionist argument explain? Suppose we make some assumptions about measures on parameter space (to define "small" changes) and also put in some appropriate initial conditions. The argument then shows that over sufficient time, in almost all the descendants of a randomly chosen universe, the parameters are nearly optimal for the production of black holes. Thus, if our universe is, in fact, almost optimal for black-hole production, then our universe is not so special after all (certainly less special than it appears to proponents of the anthropic principle). Of course, like any statistical estimate of the distribution of parameters in a population, the argument does not show why the parameters in any individual case take the precise values they do. As such, it fails Leibniz's test for explanatory power; it fails to give a sufficient reason why the universe is thus, and not otherwise. Instead, Smolin's selectionist argument may challenge the need to apply Leibniz's principle here at all, that is, the need to require an explanation for the precise values of the parameters. In the light of Smolin's model, one might be no more inclined to demand an explanation for that than to demand an explanation for the exact configuration of the stars in the sky above. Perhaps.

What of falsifiability? Falsifiability of a hypothesis depends on holding fixed the auxiliary assumptions needed to produce the targeted conclusion (otherwise, we could deflect the arrow of falsification to one of these). In practice, one tries to show that the auxiliaries are themselves well confirmed or otherwise scientifically entrenched. In Smolin's case, the target on which falsifiability rests is the claim that our universe is nearly optimal for black-hole production. As we noted above, however, that claim is not actually a consequence of the theory (only that most universes are nearly optimal). To move from this statistical conclusion to the targeted conclusion about our universe, Smolin simply assumes that our universe is typical (94, 96, 101). That auxiliary assumption, however, is neither confirmed nor entrenched. Thus, if changes in the values of our parameters did not lead to a lower rate of black-hole production, we could always "save" Smolin's theory by supposing that our universe is not typical. So the falsifiability he claims for the theory depends critically on the auxiliary, on whether we are typical among the many universes. In this context, Pierre Duhem's argument applies and there is no way to falsify the specific assumptions of Smolin's model.

One cannot help but admire Smolin's vision and his considerable achievement in melding together beautifully several different scientific programs and applying ideas from different scientific domains. His work nicely illustrates the holism to which his relationism is committed. The preceding reservations respond to one of the less revolutionary features of his work. Whereas Smolin challenges much of the received scientific wisdom of our time, he is quite conventional in his acceptance of the received methodological wisdom one finds among scientists. He accepts a sharp demarcation of science and the criteria of explanatory power and testability which are conventionally said to bound it. By his own tenets, then, if (as we suggest) those criteria do not work well for him, shall we just put his effort out to pasture as myth or mere speculation? That would be crazy. It will be clear to every reader that Smolin is not merely speculating and that his arguments are grounded in science and subject to thoughtful scientific debate and constraint. Only an impoverished philosophy of science could direct a librarian to put Smolin in with Thomas Bullfinch. It looks like Smolin needs to be liberated from the bad philosophy of science of scientists (polite indeed!) and to hang out with us philosophers a little bit more.

STEVEN WEINSTEIN
ARTHUR FINE

Northwestern University

THE JOURNAL OF PHILOSOPHY
VOLUME XCV, NO. 6, JUNE 1998

CONSENSUS, STABILITY, AND NORMATIVITY IN RAWLS'S POLITICAL LIBERALISM*

In a recent exchange with John Rawls, Jürgen Habermas[1] objects that Rawls's goal of an overlapping consensus is too weak a foundation for a theory of justice. We know that for Rawls an overlapping consensus is a merely political consensus, a consensus on a public conception of justice that does not itself admit of moral truth. The concept of moral truth is reserved for use within particular comprehensive doctrines, and from within those doctrines citizens can affirm the political conception in any way they understand as rational. These private affirmations are of no public, political use, however, since they are dependent on comprehensive doctrines that not all citizens share. So we are left in a puzzling position. On the one hand, the political conception is not justified with reference to its moral truth. Rather, it is justified by its ability to serve as the object of an overlapping consensus. But the specific affirmations that make up an overlapping consensus cannot themselves justify the political conception in a public, political way, because these depend on controversial moral doctrines that not all citizens accept. In a sense, then, Rawls's political conception is not justified by anything more than the general fact of the overlapping consensus itself. Trying to understand the nature of this fact, trying to pry open the specific affirmations that relate the political conception to moral truth, would compromise the purity of a merely political conception of justice.

What this means, argues Habermas, is that political liberalism seems to use agreement as its criterion of justification. The only

---

* An exchange with Peter Steinberger was crucial in prompting me to write this paper.

[1] "Reconciliation through the Public Use of Reason: Remarks on John Rawls's Political Liberalism," this JOURNAL, XCII, 3 (March 1995): 109-31

0022-362X/98/9506/269-92

thing that can publicly validate the political conception is the fact of the overlapping consensus, and this fact itself signifies nothing more than the citizens' separate and politically unanalyzable agreements to live by the political conception. All that seems to justify the political conception, then, is the fact that all agree to it. "When Rawls calls his conception of justice *political*," writes Habermas, "his intention appears to be...to collapse the distinction between its justified acceptability and its actual acceptance" (*ibid.*, p. 122).

The question then becomes: Can agreement serve as the basis for a theory of justice? For Habermas, the answer to this question is "no," because agreement lacks the normative content that any morally justified conception must have. For how good is agreement if what is agreed to lacks moral content? It might be useful to have agreement on our political conception, but the traditional task of the political philosopher is to identify the conception of justice that would be morally worthy of our agreement. In his recent writings, Habermas argues, Rawls seems to have abandoned this traditional project of moral justification in favor of a purely practical task. Habermas is hardly alone here. Many commentators have suggested that by emphasizing the problem of stability and in seeking a overlapping consensus, Rawls has shifted his focus from normative to purely practical or political questions.[2] By concentrating specifically on the problem of a shared political conception in the wake of the history of European religious conflict, Rawls seems to have given up the goal of moral justification in favor of securing political peace. This last project may be of considerable practical importance, but it would seem that we can and should distinguish it from the project of normative theorizing. "In my view," writes Habermas, "Rawls must make a sharper distinction between acceptability and acceptance" (*op. cit.*, p. 122).

---

[2] This view has been taken not just by Habermas and those sympathetic to Habermas's project, but also by those sympathetic to Rawls's earlier work, who regard the appeal to overlapping consensus as a change for the worse. See, for instance, Chandran Kukathas and Philip Pettit, *Rawls: A Theory of Justice and Its Critics* (Stanford: University Press, 1990), pp. 142-50; and Joseph Raz, "Facing Diversity: The Case of Epistemic Abstinence," *Philosophy and Public Affairs*, xix (1990) 3-46.

The best available response to this kind of objection is Joshua Cohen, "Moral Pluralism and Political Consensus," in David Copp, Jean Hampton, and John Roemer, eds., *The Idea of Democracy* (New York: Cambridge, 1993), pp. 270-91. Cohen's treatment of the fact of reasonable pluralism is excellent, and seems to have influenced Rawls's revisions of his later papers in *Political Liberalism* (New York: Columbia, 1992).

Yet Cohen's account is ultimately unsatisfactory, since he does not separate, as I shall do here, the different roles that the ideas of consensus and stability play in Rawls's theory. On the contrary, he sees them as belonging to the same stage of argument, introduced after the positive theory of justice has been presented. This misses the way the idea of consensus shapes Rawls's theory of justice itself, and leaves it unclear what sort of justification the second stage of argument could provide. Cohen tells us, rather vaguely, that "satisfying the pluralistic consensus condition does count in favor of the correctness of a conception of justice, and that while failure to meet it is not a sufficient reason for rejecting a conception, it would provide some reason to modify a view to bring it into conformity with that test" (pp. 273-74). Just what sort of reason Cohen does not say. Against this view, I shall argue that availability to consensus is a necessary condition of (modern) political justification, and that the argument for stability plays the separate role of showing that availability to consensus can also serve as a sufficient condition.

I want to argue that this demand for a sharper distinction is mistaken. Against Habermas and many others, I shall argue that it is not the case that Rawls has diminished the normative force of his theory of justice by emphasizing the ideas of consensus and stability, or abandoned normative theorizing in favor of a purely practical end. In my view, this criticism results from a misunderstanding of the way the ideas of stability and consensus are deployed in Rawls's theory. The critics tend to assume that they are deployed all at once, to solve a single and purely practical problem. Against this view, I shall argue that they are deployed in a serial fashion, in a way that reflects a developing normative attitude toward that practical problem. There is nothing wrong with the claim that the notions of consensus and stability have moral content in themselves, provided that we introduce these notions in the right way.

I CONSENSUS AND NORMATIVITY

There is a sense, though perhaps a seemingly trivial one, in which any normative theory aims at consensus. A normative theory is an account of the way (or at least a way) it would be good for us to act. But clearly our acting in a particular way depends on our accepting particular beliefs. Having claimed that something is good for us to do, then, the normative theorist also must hold that there ought to be a consensus around certain beliefs, specifically those which would produce good actions under ordinary conditions.

Of course, these beliefs are not always identical to those held by the normative theorist, as even a brief glance at the history of moral and political philosophy can show. For instance, the lower classes of Plato's *kallipolis* do not believe in the existence of Forms, and many utilitarians do not want ordinary agents to rely on the principle of utility in everyday moral judgment. These are especially interesting examples, and I shall continue to refer to them as I proceed. In these cases, the normative theorists distinguish the moral beliefs that are fully justified by the theories from those which would produce good actions under ordinary conditions. Still, in specifying that second set of beliefs, these theorists do specify beliefs that it would be good for everyone to have under ordinary conditions. I shall call the consensus around these favored beliefs a *normative consensus*.

This talk of normative consensus seems trivial, because it seems to specify nothing about the content of that consensus. But we can extract from it a substantive, though extremely minimal, normative claim. The claim is just that it would be better, other things being equal, to have a normative consensus than not to have one. The normative theorist's emphasis on action plays a crucial role here. It is not clear that an epistemic claim, a claim about what is true, necessarily implies a claim about what it would be good for everyone to believe. Perhaps there is some reason why it would be better if at least some people were ignorant of what is actually true. In the epistemic case, we can separate (at least potentially) the question of what the theorist identifies as true from the question of what others should believe. In the practical case, however, we cannot completely separate what the theorist identifies as morally justified from what others should believe. The normative theorist can still recommend that others should have *different* beliefs, but she has to recommend that they have some beliefs, namely, the ones that would prompt good actions; and she must hold that it is a good thing for them to hold those beliefs. In that sense, a consensus around the beliefs is a normative improvement over a state of affairs in which there is no such consensus. Without this claim, the theory would not be normative: it would not guide our actions.

This minimal claim may seem more forceful when it is understood in negative terms. If we can say that a normative consensus itself represents a normative improvement over the state of affairs in which there is no such consensus, we can also say that a state of affairs in which such a consensus is not yet achieved is a worse state of affairs than one in which the consensus has been achieved. Even more strongly, we can say that, if there exists some fact that makes it impossible, or at least extremely unlikely, that a normative consensus can be achieved, that fact must itself be understood as bad from within the normative theory itself. If we call such a fact a *consensus-hindering fact*, we can say that any normative theory must regard a consensus-hindering fact as a bad thing, an obstacle that ought to be addressed by the theory itself.

How might a normative theory address a consensus-hindering fact? On the one hand, it might direct those who hold the theory to take measures designed to change the fact. These measures would not themselves be the good actions originally specified by the theory. But as steps designed to remove the consensus-hindering fact as an obstacle to those actions, the measures would have to be identified by the theory as instrumentally good.

Suppose, on the other hand, that such measures are extremely difficult, or perhaps even impossible, to achieve. Suppose that the consensus-hindering fact is a *stubborn fact*, a fact that cannot easily be changed. In that case, the normative theory would itself be stubborn or foolish to direct its adherents to change or overcome the consensus-hindering fact. A *stubborn consensus-hindering fact* is one that adherents of the normative theory can only cope with, by adjusting their recommendations to the reality of the fact. The theory could do this by recommending measures to minimize the effect of the fact on the achievement of a normative consensus. Such measures would once again have to be recognized by the theory as secondarily good actions. But in this case, the secondarily good actions would be understood not as instrumental to the achievement of primarily good actions, but as second-best measures designed to maximize the achievement of primarily good actions under actual conditions.

The key point here is that a normative theory will naturally be reluctant to recognize a consensus-hindering fact as stubborn. It would rather take the steps designed to achieve a normative consensus around the beliefs it identifies as favored. If, however, such a consensus is difficult or impossible to achieve, even through additional or instrumental measures, the theory will have to modify its aspirations to cope with the facts of the world. In such a case, the theory may even have to modify the content of the normative consensus, the content of the beliefs it primarily identifies as favored, in order to achieve the actions it primarily identifies as good.

We can see these kinds of modifications at work in the examples I have already mentioned, namely, Plato's *kallipolis* and certain versions of utilitarianism. In the *Republic*, Plato recognizes that it will be extremely difficult, if not impossible, to achieve a normative consensus around the idea that only philosophers ought to rule.[8] The ordinary citizens of Athens, Plato believes, accept the mistaken belief that they are entitled to participate in the democratic governance of the polis. To overcome this fact, Plato recommends an elaborate set of educational reforms meant to socialize the citizens of the *kallipo-*

---

[8] Here, I simply rely on Plato's account of the *kallipolis* for purposes of example; I make no argument against those who claim that Plato did not really understand this city as an ideal. My own view is that such claims are unconvincing, but there is no need to address them here.

ls into holding the belief that only the members of the highest class are entitled to rule. Education, for Plato, is an instrumental good designed not just to create philosopher rulers, but also to create a normative consensus around the idea that philosophers ought to rule. Yet even the best Platonic education cannot convince ordinary citizens that philosophers ought to rule because they are philosophers. Convincing the public that philosophy aims at real knowledge rather than idle wordplay is for Plato an impossible task: there is no way that ordinary, cave-dwelling citizens can or should recognize the existence of Forms as the true objects of knowledge. The ordinary person's understanding of philosophy is not just a consensus-hindering fact but also a stubborn fact, one that cannot be overcome through additional education or anything else. Hence Plato recommends that his educational and political program include a number of myths and lies designed to reconcile ordinary citizens to the fact of philosophical rule. Since Plato regards knowledge or truth as the highest good, it cannot be that these myths and lies are themselves primarily good; nor are they even instrumentally good, as they do not promote knowledge or truth. Rather, they are second-best measures designed to allow Plato to cope with a stubborn consensus-hindering fact. Ordinary citizens can be brought to believe that only members of the highest class ought to rule. But since they cannot be brought to believe that members of the highest class ought to rule because they are philosophers, they must be brought to believe that it is the gold in their veins that entitles the philosophers to rule.

In the case of utilitarianism, the relevant fact is the difficulty that ordinary agents have in performing utilitarian calculations under ordinary conditions, or perhaps the difficulty of understanding the relation of particular actions to the general utility. To address these facts, utilitarians have tended, like Plato, to stress the importance of education, particularly education about the workings of large-scale social institutions. Such education, utilitarians like J. S. Mill suggest, will bring ordinary agents to appreciate the benefits that general social arrangements provide to ordinary agents like themselves. Such appreciation would, in turn, make it more likely that the ordinary agents would act as the utilitarian calculus dictates, to do the things that would promote and maintain social institutions that benefit the mass of ordinary citizens. By closing the gap between individual motivation and large-scale social outcomes, education is instrumental in promoting a normative consensus around utilitarian principles. Still, it may turn out to be that full-scale utilitarian calculations are too difficult even for well-educated individuals. In such a case, the utilitarian may recommend an education that brings about a deep and even overriding commitment to certain simpler and subordinate principles of actions. Such a strong commitment to subordinate principles would itself be in conflict with the principle of utility, but a utilitarian might nonetheless seek to promote such a commitment to deal with the stubborn and intractable limitations of individual reasoners. In such a case, utilitarian education would seek a normative consensus not around utilitarianism, but around the simpler and subordinate principles.

Notice that in these cases, the normative theories are explicitly concerned with the extent and limits of individual rationality. This should not be surprising. The theories seek a normative consensus around the beliefs they identify as justified. If that consensus is difficult to achieve, they can only understand others' failure to join in the consensus as a failure of rationality, as a refusal to accept what has already been justified. Hence the theories recommend education as an instrumental measure to overcome the obstacle of limited rationality. But if this obstacle is too difficult to overcome, if full education is impossible or unworkable, the theories recommend a kind of second-best normative consensus, a consensus around beliefs that are less than fully rational. In such cases, the theories make explicit concessions to the limits of individual rationality.

Suppose, however, that there exists a stubborn and consensus-hindering fact that is ascribable not to the limits of individual rationality, but to the workings of reason itself. Suppose, that is, that individuals exercising their rational capacities in a full and undiminished manner produce a state of affairs that makes it impossible, or nearly impossible, to achieve a normative consensus around a particular set of beliefs. If a normative theory then came to identify one or more of those beliefs as fully justified, the normative theory would be faced with a grave difficulty. As a normative theory, a theory meant to guide our actions, it would seek to secure a normative consensus around the beliefs it identified as justified. But in this case, by hypothesis, such a normative consensus would be impossible or nearly impossible to achieve, even by instrumental measures. If the theory sought to achieve a normative consensus around a second-best set of beliefs, it could no longer justify this move as a necessary concession to the limits of individual rationality. Rather, the normative theory would be asking individuals to suspend the workings of their rational capacities for the sake of considerations that another exercise of reason identified as justified. We would be faced, then, with a conflict within reason, one that the normative theory could not settle without compromising its own claim to rationality.

Putting all of these points together, we can construct the following argument relating consensus to normativity. It is crucial to see that this argument is not a pragmatic argument that depends on our being motivated to solve any practical problem. The argument does not depend on an interest in achieving consensus extrinsic to the project of normative theorizing itself. Rather, the argument is a deductive argument from the very idea of a normative theory. The argument runs as follows. For any normative theory, if it is the case that:

(1) there exists a consensus-hindering fact;
(2) that fact is stubborn, that is, impossible or nearly impossible to overcome;
(3) the fact is ascribable not to the limitation, but to the full exercise of individual rationality;
(4) there is not to be a conflict within reason;

then we may conclude that:

(C) the normative theory must be such that the beliefs it identifies as fully justified do not conflict or interfere with the consensus-hindering fact.

I take my discussion to this point to have established that this argument is valid.

II REASONABLE PLURALISM AND REASONABLE POLITICAL JUSTIFICATION

With this argument in mind, we can turn back to Rawls, and specifically to what he calls the *fact of reasonable pluralism*. The recognition of this fact is the recognition, after reflection on the history of European religious conflict, that modern, pluralistic societies will almost certainly not be able to achieve a normative consensus on the ultimate aims of life, on the values that are the concern of traditional religious doctrines. In accordance with the argument I have just developed, I shall understand the public disagreement implied by the fact of reasonable pluralism not as a practical problem in need of a solution, but as a consensus-hindering fact that satisfies not just (1) but also the additional premises (2)-(4) above, and thus as a fact that alters the very nature of normative theorizing under modern conditions.

I shall not spend any real time here arguing for (1) and (2); that is, I shall not provide an argument for the claims that modern moral and religious pluralism is an obstacle to public consensus, and that this obstacle is a stubborn one, one that is not likely to be overcome by anyone's efforts.[4] I take these claims to be simply empirical premises, already adequately justified on historical and sociological grounds, and I take it that Rawls understands them as justified in the same way (*op. cit.*, pp. xxi-xxviii). Notice that I do not stop to list or even to categorize the specific claims that are now no longer available to public consensus under conditions of modern moral and religious pluralism. For purposes of example, the question of the status of Biblical morality, and the closely related questions of who is to decide what Biblical morality is and by what interpretive means, can stand as a kind of marker for the public disagreement Rawls has in mind. At this stage in the argument, it is not crucial that we specify the exact nature of the claims that the fact of reasonable pluralism puts into public dispute. (Indeed, at *no* stage of the argument will it be important that we specify the nature of these claims, as I shall emphasize below.) All we need to know is that it does put into public dispute some questions about which governments have sought to legislate, and about which they might continue to legislate. This is enough for us to understand the fact of modern pluralism as a stubborn and consensus-hindering fact for purposes of normative political theorizing. That this is a stubborn and consensus-hindering fact is all we need to know; the extent of this fact is another matter, one that is irrelevant to establishing the truth of (1) and (2).

Premise (3), however, is not simply a historical or sociological claim, since it requires that we see the fact of modern pluralism as itself reasonable. We must understand the diversity of moral and religious doctrines not as a failure of rationality, but as a consequence of rationality in its diverse exercises. This is a very strong claim, and many have minimized its importance in Rawls's argument. The more typical reaction is to stress how much Rawls demands of traditional religious believers, namely, that they suspend their own doctrinal commitments for purposes of political justification. It is easy to miss how much Rawls demands of secular liberals, namely, that they understand the diversity of religious traditions as themselves expressions of reason. This does not mean that secular liberals must recognize religious doctrines as true, or even as rational within their own secular terms. But it does mean that secular liberals must understand the various surviving religious traditions as coherent bodies of thinking,[5] as rational within their own religious terms, and thus as sincere and disciplined efforts to understand the meaning and purpose of human life.[6]

---

[4] "... the diversity of reasonable comprehensive religious, philosophical, and moral doctrines found in modern democratic societies is not a mere historical condition that may soon pass away; it is a permanent feature of the public culture of democracy. Under the political and social conditions secured by the basic rights and liberties of free institutions, a diversity of conflicting and irreconcilable—and what's more, reasonable—comprehensive doctrines will come about and persist if such diversity does not already obtain" (Rawls, p. 36).

Here, it is not enough to see others' comprehensive doctrines as impressively coherent and systematic. It is possible to admire the disciplined development of a doctrine while nonetheless regarding the doctrine as developed precisely to cover over and thus to apologize for an exercise of power and domination. For instance, we might admire the resourceful efforts of antebellum Southern theologians to justify slavery while nonetheless regarding their efforts as fundamentally ideological, as specifically designed to apologize for a vicious social practice. In such a case, we would regard the developed doctrine as an inauthentic use of reason, an exercise of reason designed to serve an end that is itself opposed to reason. Rawls insists on something stronger than this: that we see others' comprehensive doctrines not as ideological or apologetic, but as authentic exercises of reason to comprehend the way human beings should live.[7] This requires that we see the differences that prevent us from affirming their comprehensive doctrines not as the product of prejudice or self-deception, but as the consequences of what Rawls calls the *burdens of judgment* (*op. cit.*, pp. 54-58). By the burdens of judgment Rawls means the factors that may lead even sincere and disciplined reasoners to disagree: the complexity of evidence, the indeterminacy of abstract concepts, our different experiences, the relative weight we give to relevant considerations. We must be able to look at other comprehensive doctrines and to say: had we some other experience and thus some other way of looking at the world, we might be able to affirm these doctrines as our own, without bias or self-deception. This should not be taken for a positive claim, a claim that we understand what it would be like to affirm the other doctrines. Rather, it is a merely negative claim, a claim that we have no reason to regard the coherence and systematicity of the other doctrines as anything but the sincere and disciplined application of reason to the ends and values of some other kinds of life.

---

[5] "...a reasonable doctrine is an exercise of theoretical reason: it covers the major religious, philosophical, and moral aspects of human life in a more or less consistent and coherent manner. It organizes and characterizes recognized values so that they are compatible with one another and express an intelligible view of the world.... In singling out which values to count as especially significant and how to balance them when they conflict is also an exercise of practical reason.... Finally, a third feature is that while a reasonable comprehensive view is not necessarily fixed and unchanging, it normally belongs to, or draws upon, a tradition of thought and doctrine...it tends to evolve slowly in the light of what, from its point of view, it sees as good and sufficient reasons" (Rawls, p. 59).

[6] "Religious and philosophical doctrines express views of the world and of our life with one another, severally and collectively, as a whole" (Rawls, p. 58).

[7] "[Reasonable comprehensive doctrines] are not simply the upshot of self- and class interests, or of people's understandable tendency to view the political world from a limited standpoint. Instead, they are in part the work of free practical reason within the framework of free institutions" (Rawls, p. 37).

The positive claim that does follow from these negative understandings of other comprehensive doctrines is this: that we regard the diversity of comprehensive doctrines that exist under modern conditions as "the natural outcome of the activities of human reason under enduring free institutions" (*op. cit.*, p. xxiv). Here, we need not specify the meaning of "free institutions" beyond the thought of institutions that leave the various comprehensive doctrines free of harassment or persecution, and thus free to develop, express, and modify those doctrines in ways that adherents or potential adherents of the doctrines might find convincing. Since comprehensive doctrines are exercises of reason, we need only be concerned here with the freedom to engage in rational activity. The positive claim that Rawls insists on is just that, given the diversity and complexity of modern societies, different persons reasoning from different points of view will, if left alone, produce a variety of comprehensive doctrines, none of which will prove universally persuasive. The inability to achieve consensus results not from a failure to reason, or from a failure of reason, but from the natural exercise of reason under conditions of pluralism. This is what is required to establish premise (3).[8]

---

[8] Leif Wenar objects that the fact of reasonable pluralism and the account of the burdens of judgment should never have been included in Rawls's argument, since they diminish rather than increase the likelihood of achieving an overlapping consensus around justice as fairness—"*Political Liberalism*: An Internal Critique," *Ethics*, CVI (October 1995): 32-62. Wenar's main argument for this is an appeal to Catholic doctrine, which (as Wenar presents it) regards the post-Reformation splintering of Western religious thinking as unfortunate if not disastrous. On this view, says Wenar, the existence of other, non-Catholic comprehensive doctrines is the consequence not of the free exercise of human reason and the burdens of judgment, but of human weakness and depravity. Since Rawls says that his theory can be affirmed by all the major historical faiths, presumably including Catholicism, Wenar concludes that this part of Rawls's argument should have been dropped to facilitate the achievement of an overlapping consensus.

This argument is deeply problematic, however. Wenar assumes exactly the reading I am trying to oppose here, that is, that Rawls's main goal is to achieve a consensus among actual comprehensive doctrines. But this is not the case: Rawls is trying instead to show, through philosophical argument, why availability to consensus is a necessary condition of (modern) political justification. Without the accounts of the fact of reasonable pluralism and the burdens of judgment, it is unclear why a consensus that includes or excludes Catholics or anyone else should be of any interest to normative theory. Without these features of Rawls's

Premise (4), however, requires a rather different treatment. Premises (1)-(3) are all descriptive claims: the first two straightforwardly historical or sociological, the third a more complex judgment about the effects of rational activity under certain historical and sociological conditions. Premise (4), if we accept it, is not a descriptive claim at all, but a determination to see the results of normative political theorizing as avoiding a conflict within reason itself. Here, we specify nothing about the content of our normative political theory. We say only that we want to see that content as rational, and that this requires that this content not conflict with what other exercises of rationality identify as justified. We say only that potential conflicts in reason are to be avoided if at all possible. But since the diversity of comprehensive doctrines is itself the product of rational activity, we arrive at the conclusion of the argument, namely, that the content of our normative political theory must be such that it does not conflict with the diversity of comprehensive doctrines. This conclusion just is the claim that a modern political theory must be reasonable, that we must be able to justify it to all citizens in terms that each of them can accept.

It is important to see that at this point in the argument, the reasonable is simply a necessary condition that any modern political theory must meet if it is to avoid a conflict in reason. One could imagine specifying the content of that political theory in any number of ways: so long as the requirement of reasonableness was met, we might go on to include any number of positive normative claims based on any number of considerations we understood as rational.

---

argument, the theory is vulnerable to Habermas's objection, even if Catholics happen to affirm it.

Still, if we allow the account of reasonable pluralism that Wenar says we should drop, should we not be concerned that Catholicism does not affirm it? Well, only if Catholicism *could not* affirm it, and only if Catholicism is a reasonable comprehensive doctrine. Wenar establishes neither of these. In support of the first, he merely quotes Catholic thinkers who *do not* affirm it; and in support of the second, he suggests only that their Catholic doctrine affirms something quite like Rawls's two principles of justice. Here, again, Wenar is wrongly fixated on the practical goal of achieving consensus on the results of Rawls's theory. But Rawls is concerned not simply with results, but with showing that those results have rational justification. So his goal is not so much the two principles, but rather showing that the two principles follow, via the original position, from the idea of reasonable political justification. Since Wenar never shows that his Catholic thinkers accept the idea of reasonable political justification along with their pessimistic account of modernity, he never shows that their particular Catholicism is a reasonable comprehensive doctrine. In all of this, Rawls's hypothesis of a reasonable Catholicism which could affirm both political liberalism and the doctrine of the "One True Church" remains unrefuted.

To say that under modern conditions a political theory must be reasonable in order to be rational is not yet to say that it might not also be rational in certain other ways. Indeed, one might suppose that a political theory would have to include much more than the criterion of reasonableness in order to be understood as fully justified. The striking feature of Rawls's political liberalism, however, is that it rejects this thought. Rawls proposes instead that we build our entire political theory on the criterion of reasonableness, on the thought that political arrangements ought to be justified in ways that all can accept.

But why might one make such a proposal? It does not follow from the argument I have defended as valid in the preceding section, and sound in the present section. That argument tells us only that, if we are to avoid a conflict in reason, we must avoid the stubborn and consensus-hindering fact of modern pluralism. We must justify our political claims on terms that all can accept, no matter what (reasonable) comprehensive doctrines they affirm as rational. But that is not yet to say that this necessary condition of political justification should also serve as a sufficient condition. So what is the source of this additional suggestion?

Recall that I have not specified any account of the particular claims that the modern plurality of comprehensive doctrines has put into public dispute. I have said only that there are such claims, and that these are likely to serve as a stubborn obstacle to normative consensus. Suppose, then, that we wanted to include in our political theory a requirement of rationality that extended beyond the requirement of the reasonable. Suppose, that is, that this additional requirement identified certain political claims as justified, and that these claims were not themselves implied by the notion of the reasonable. Because the requirement of reasonableness remains in place, we would have to check the new, supposedly rational claims for reasonableness; that is, we would have to determine whether the new claims were in conflict with those of any particular (reasonable) comprehensive doctrine. But this would require that we look into the details of particular comprehensive doctrines to see what they identify as rational, and then to see if their claims conflict with what our political theory had identified as rational. But this is precisely what we have not yet done. We have not looked at the details of particular comprehensive doctrines, seeing what they affirm and what they do not. We have said only that they affirm some things and not others, and that this is likely to be an obstacle to public consensus. The persistent possibility of public disagreement is enough to motivate the bare requirement of reasonableness. But if we added to this requirement a further criterion of rationality, we would immediately be involved in assessing this criterion against all the criteria of the specific comprehensive doctrines. We would have to move from noting the persistent possibility of public disagreement to a detailed investigation of particular doctrines and the claims they identify as justified.

Given the diversity of comprehensive doctrines, this kind of investigation is likely to be arduous and ultimately inconclusive. How can we assume, from the standpoint of particular individuals holding particular comprehensive doctrines, that we have fully understood the details of all other comprehensive doctrines? At what point can we say that every available comprehensive doctrine has been evaluated, and found to affirm (or at least not to deny) what our political theory has identified as rational? All we can ever say is that we have investigated all the comprehensive doctrines familiar to us. But this makes political justification hostage to existing circumstance, to existing comprehensive doctrines, in exactly the way that Habermas fears. Hence we do not "look at the various comprehensive doctrines actually found in society and specify an index of such goods likely to be near those doctrines' center of gravity, so to speak.... This is not how justice as fairness proceeds; to do so would make it political in the wrong way" (*op. cit.*, pp. 39-40).

Instead, we treat the idea of the reasonable as a freestanding source of political justification. On its own, the requirement of reasonableness swings free of every comprehensive doctrine. But coupled with any further requirement of justification, the requirement of reasonableness itself demands a detailed and finally question-begging investigation of what different comprehensive doctrines affirm. To avoid this kind of investigation, to avoid specifying what particular doctrines affirm, we have to stick to the requirement of reasonableness itself. By allowing the reasonable to stand as the sole criterion of political justification, we are able to leave all comprehensive doctrines to the side. We do not have to determine whether our political claims are justified to each person, given the details of her comprehensive doctrine. Rather, we take the fact that our political claims are in principle available to any person to count as our standard of political justification. As a necessary condition of modern political justification, consensus seems to demand a problematic inquiry into the sources of conflict and disagreement. But if we allow consensus to serve as also a sufficient condition of political justification, we can leave these sources to the side; for the possibility of agreement is already built into reasonable political justification itself. The only clear way to ensure that our political claims are reasonable is to make reasonableness the only standard by which we judge our political claims.[9]

III. STABILITY AND NORMATIVITY

Two problems remain. The first is that of content: we do not yet know whether the bare idea of reasonableness is enough to justify anything like a set of principles or institutions for an actual society. Rawls's answer to this problem, of course, is his positive theory of justice. He tries to show how the notion of the reasonable is captured by the description of the parties in the original position, and then, derivatively, by the choice of those parties for the two principles of justice. Like many others, I have serious doubts about each of these stretches of argument, even though I am largely sympathetic to Rawls's project. I shall not pursue these doubts here, since my present interest is not in what political liberalism can justify, but in the way it chooses to justify its claims. So I shall leave the specific content of the theory aside. I shall simply assume, perhaps optimistically, that the notion of the reasonable can itself be specified in a way that implies some positive content for a theory of justice.[10] I shall turn instead to the second problem that remains for Rawls's political liberalism, namely, that of stability. Having identified certain political principles as justified, Rawls then goes on to ask whether actual individuals could consistently affirm and comply with these principles under actual conditions.

Like that of consensus, this notion of stability or feasibility might seem extrinsic to the traditional project of normative theorizing. Surely, a normative theory wants its adherents to live up to the principles it identifies as justified; and surely, it would be a matter of concern if living up to those principles turned out to be difficult or impossible. But this would seem to be a problem not for the normative theory itself, but for the people who could not live up to its principles. Or, if the normative theory did address the problem, it would address it not by modifying the principles it identified as justified, but by offering the public some second-best set of principles that would best approximate the effects of the justified principles under actual conditions.

---

[9] This conversion of a necessary into a sufficient condition of justification is a definitively Kantian move. In Immanuel Kant's practical philosophy, the moral requirement of universalizability is not derived from practical reason itself. Rather, conformity to universal law is simply a necessary condition that any principle of practical reason would have to meet. Instead of deriving some principle of practical reason (which would then have to be shown to be universally valid), Kant proposes that we allow the formal requirement of universality to serve as our principle of practical reason, thus converting a necessary into a sufficient condition of justification. The reason we allow this is, I think, quite like Rawls's own: because we want to have some way of rationally justifying our actions to all others. I attempt to make the case for this understanding of Kant in my "What Kind of Law Can This Be? Kant's Derivations of the Categorical Imperative" (unpublished manuscript).

[10] I formulate my own criticisms of Rawls's arguments, and my own alternative for deriving political content from the notion of the reasonable, in my *Kant and the Possibility of Politics* (unpublished manuscript). For a much-compressed account, see my "Formal Liberalism and the Justice of Publicity," *Proceedings of the Eighth International Kant Congress*, II (March 1995): 61-69.

Like Plato's and Mill's treatments of consensus, however, this way of thinking assumes that a failure of stability is a failure of rationality—this time not a failure to recognize what reason can justify, but a failure to act on what reason has justified. But Rawls's concern with stability, like his concern with consensus, is internal to reason and thus internal to the project of normative theorizing itself. In the case of consensus, Rawls is concerned not simply with disagreement, but with disagreement as the result of rational activity under modern conditions. He then extracts the notion of the reasonable as a necessary condition of a justified normative theory in our time. Now, in the case of stability, Rawls is concerned not simply with what persons can live up to, but with what they can live up to as rational beings. The specific question of stability that Rawls poses is whether rational agents can affirm the claims of political liberalism as fully justified.[11]

This problem arises because, as we have seen, the notion of the reasonable is not itself a sufficient condition of political justification. It is only a necessary condition of political justification under conditions of reasonable pluralism. But we go on to allow this necessary condition to serve as also a sufficient condition so that we may avoid a detailed examination of particular comprehensive doctrines. It is this move, and this move only, that makes the problem of stability central to Rawls's theory of justice. According to more traditional theorists like Plato and Mill, political claims derive their authority from reason itself. For Plato, the laws get their force from the philosopher ruler's insight into Forms, and Forms are the necessary objects of rational activity. For Mill, justice is a device to promote utility, and utility is the natural object of practical reason. In both cases, political justification is simply rational justification. In Rawls's case, however, political justification is reasonable justification—which we allow to serve as a political proxy for rational justification. The question is now how we can allow this, given what we take to be rational.

---

[11] The only reading of Rawls along these lines seems to be that of Thomas E. Hill, Jr., "The Stability Problem in Political Liberalism," *Pacific Philosophical Quarterly*, DCCV (1994): 333-52. Hill stresses, as I do here, that the concern with stability is less a matter of practical survival than of rational justification. He does not distinguish as sharply as I would prefer between necessary-and sufficient conditions of political justification, but I take his reading to be consistent with my own.

Since what we take to be rational is a matter for our own comprehensive doctrines, the problem of stability is ultimately to be addressed not by the normative theory, but by the comprehensive doctrines themselves. Rawls did not always appreciate this. In *A Theory of Justice*[12] the problem of stability was addressed by an appeal to a minimal, instrumental sense of rationality (assumed to be common to all richer conceptions of rationality) and to certain minimal psychological assumptions (most prominently the Aristotelian Principle, assumed to hold of anyone endowed with rational capacities) (*ibid.*, part III). In the terms of his later work, it was as if Rawls had searched the various comprehensive doctrines for a lowest common denominator of rationality; he then attempted to use this lowest common denominator to prove that the reasonable was rational on every comprehensive doctrine. But the guiding thought of the later work, and especially of the idea of the reasonable, is that this sort of procedure is both unwieldy and unnecessary. It is unwieldy because it would involve us in an elaborate survey of comprehensive doctrines to determine if our minimal conception of rationality was common to them all; and it is unnecessary because the particular comprehensive doctrines, themselves exercises of reason, are perfectly capable of affirming the reasonable in their own terms. So long as these various private exercises of reason converge on the same public notion of the reasonable, there is no reason for the normative theory itself to secure the stability of a political conception of justice.

What the normative theory does instead is to lay out the idea of an overlapping consensus, and to explain why this convergence of various comprehensive doctrines is sufficient for the purposes of the theory. Beyond this, the overlapping consensus is or is not a historical fact, a fact that the theory can hypothesize but that only private citizens and their particular doctrines can affirm.[13] All Rawls can do is to suggest certain ways that comprehensive doctrines might come to affirm the reasonable as a basis for political justification. He does this by laying out what he calls a "model case" of three sorts of persons holding three possible kinds of comprehensive doctrines.[14] The first sort of person is what Rawls calls a *comprehensive liberal*, a person committed to a modern, secular moral theory of the kind defended by Immanuel Kant or Mill. The second sort of person is what Rawls calls a *pluralist*, a person who sees no need to systematize potentially conflicting values into an organized whole. The third sort of person is a *traditional religious believer*, one who holds that the highest values are revealed to us by God.

---

[12] Cambridge: Harvard, 1971.

[13] I take this notion of the overlapping consensus as a having the status of a hypothesis from Samuel Freeman.

[14] *Political Liberalism*, pp. 145-49, 168-72

These sorts of persons are chosen not simply because they are familiar to us, but more importantly because they represent the range of options open to the comprehensive doctrines as they seek to affirm the reasonable. Recall that the reasonable is a necessary but not yet a sufficient condition of modern political justification. Each comprehensive doctrine must itself defend a conception of rationality that vindicates the reasonable as sufficient for political justification; and there are, in fact, only three ways in which this might be done.

First, a comprehensive doctrine might defend a conception of rationality that exactly matches the values of the reasonable. This is what the comprehensive liberalisms do. They derive the political autonomy of reasonable citizens from the moral truths expressed in one of the prominent secular moral theories of modernity, from moral autonomy as defended by Kant or from utilitarianism as defended by Jeremy Bentham, Mill, and their successors. In the case of Kant, the requirement of reasonable political justification follows closely from the nature of practical rationality itself: to be a practically rational being, for Kant, is just to accept a principle that one's own actions be at least potentially acceptable to all others. What is practically rational, on this view, is just what could be justified to all rational beings. For utilitarians like Mill, on the other hand, reasonable political justification follows not from the nature of practical rationality, whose essential concern is utility, but from the institutional and informational conditions necessary to realize the object of practical rationality, to satisfy the desires of a large citizenry under actual conditions. For the utilitarian, reasonable political justification is a way of ensuring that all persons can understand, affirm, and comply with liberal principles—principles that, in turn, ensure that citizens have the freedom, the knowledge, and the ability to satisfy their desires in a more efficient way.[15] But even on this instrumental defense, liberal political values have their origin in moral values, rationally understood. The political autonomy of reasonable citizens follows from the proper account of practical rationality, on both a Kantian and a utilitarian view. For these views, what is reasonable just is what is rational.

---

[15] *Ibid.*, p 170. This utilitarian argument has two parts. First, there is the claim that a scheme of liberal rights is institutionally suited to promote utility on a large scale. But to this claim there must be added the further claim that a liberal regime is best ensured by a citizenry that justifies its political claims via the idea of the reasonable rather than the principle of utility itself. Here, the utilitarian will invoke the usual premises about the limits of individual knowledge, the difficulty of per-

Second, a comprehensive doctrine might defend a conception of rationality thinner than that of the reasonable. This is what pluralism does. It is a fundamentally negative doctrine, holding that values need not be organized according to a unified criterion of rationality. Pluralism is then free to affirm the value of political agreement in its own terms, as providing for social cooperation and peaceful interaction. These goods are of special importance for the pluralist, who is impressed by the diversity of human ends and by our inability to organize them into a coherent whole. Given these facts, the pluralist is in a position not only to appreciate the value of tolerant political arrangements, but also to give this value great weight against other values, and thus to affirm the binding force of reasonable political authority. This does not mean, however, that the pluralist justifies liberal political arrangements as best suited to the simultaneous pursuit of the variety of human ends. This kind of view may be perfectly reasonable, but in its instrumental derivation of liberalism from the essential diversity and incompatibility of values, it is closer to the comprehensive liberalism of Mill than to pluralism as described by Rawls. The Rawlsian pluralist says not that liberalism follows from pluralism, but only that liberal political values, understood in their own terms and appreciated for their own sake, are consistent with pluralism, even when they are given great weight. Pluralism is thus invoked not to derive liberal values, but only to justify the legitimacy of defending liberal values in their own terms.[16] On this view, the reasonable is just one sort of value we can affirm as rational.

---

forming utilitarian calculations in particular cases, and so on. (See *ibid.*, p. 162, and the discussion of utilitarianism in the first section of this paper.)

The second premise is necessary to the argument, because Rawls wants the overlapping consensus to be a consensus not just on policies or even principles of justice, but on a framework of political justification. To be included in the overlapping consensus, the utilitarian must affirm not just the rationality of liberal principles, but also the rationality of citizens' deriving liberal principles from the idea of the reasonable.

[16] The difference between the two types of pluralist views might be illustrated in the following way. The non-Rawlsian pluralist—whom we might call a *rationalist pluralist*—understands liberal toleration as valuable because it conduces to our many ends. She reasons that since we all have a great many ends, we need a liberal regime to protect all these possible sources of value. On this view, given the truth of pluralism and the existence of the diversity of human ends, we are rationally committed to reasonable political justification.

Finally, a comprehensive doctrine might defend a conception of rationality thicker than that of the reasonable. This is what the religious doctrines do. They hold that true value is religious or spiritual. But if they also hold that religious or spiritual values cannot be authentically transmitted by authority or coercion—what Rawls calls the doctrine of "free faith"—they can also affirm the value of reasonable political justification.[17] Precisely because believers must accept religious values freely, there is no justification for deriving political values from religious doctrines. Instead, we need a freestanding source of political justification, one that can distinguish itself from religious value but that can also leave ample room for the appreciation of religious value in our lives. The idea of the reasonable—which affirms only the value of shared political justification and says nothing about any religious value—is perfectly suited to this purpose. For the religious believer, politics is not of any ultimate value; but for that very reason, politics must swing free of the sources of ultimate value. The religious doctrines thus affirm the reasonable not as rational in itself, but as a rational accommodation to a part of life that we should expect to be resistant to rationality.

The Rawlsian pluralist stops short of this claim. Given only that there are a great many incompatible goods, she does not want to say every person, no matter what ends she has, is rationally committed to protecting the ends of others or even the range of her own ends. A pluralist might well value these things, but this valuation does not derive from pluralism itself. All pluralism says is that there are a great many possible sources of value which it is neither possible nor necessary to evaluate in a rational way One could be committed to this view while nonetheless pursuing a solitary end that gave no weight to social cooperation or to the ends of others.

Rawls is not concerned with ruling such a person out. Rather, he is concerned to show that a different sort of person could use pluralism to validate her commitment to reasonable political justification. Given the truth of pluralism, it is possible to regard social cooperation and common political arrangements as good in themselves, without further justification. Given the diversity of human ends, it is possible to give this value great weight against other values. While the rationalist pluralist wants to deduce a reasonable political conception from pluralism itself, the Rawlsian pluralist is concerned only to use pluralism to defend the possibility of our valuing a reasonable political conception for its own sake.

The best-known pluralist defense of liberalism, that offered by Isaiah Berlin, seems closer to the rationalist view—see "The Pursuit of the Ideal," in *The Crooked Timber of Humanity* (New York: Knopf, 1991), and "Two Concepts of Liberty," in *Four Essays on Liberty* (New York: Oxford, 1969). Rawlsian pluralism might be better represented by the kind of view presented in Michael Walzer's *Spheres of Justice* (New York: Basic, 1983). It should be noted, however, that Berlin does not try to derive liberal values directly from the truth of pluralism; rather, he appeals to certain additional premises, such as our social and communicative natures. I am indebted to Claire Curtis for clarification on these points.

[17] See *Political Liberalism*, p. 145, especially the footnote about John Locke's account of toleration. It is, of course, an open question whether the main religious doctrines of our time are in fact committed to this account. See *ibid.*, p. 170.

What the model case is meant to show is that we can have sufficient reason for pursuing reasonable political justification, whether our conception of rationality is thinner than, thicker than, or identically tailored to the boundaries of the reasonable. In each case, the comprehensive doctrines come to affirm the rationality of reasonable political justification. They solve the problem of stability by deploying their fullest conceptions of rationality, not by appealing pragmatically to the value of peaceful and stable political institutions. Only the pluralist even speaks of such pragmatic value, and even she does not appeal to pragmatism alone. Rather, the pluralist deploys a kind of (negative) philosophical argument defending the appeal to pragmatism as rationally sufficient. In no case is stability simply a matter of practical accommodation.[18]

Stability is also not a matter of settling for a kind of justification that is in any sense weak or second best. For the comprehensive liberal, the reasonable can be rationally defended in the fullest sense; for the pluralist, the reasonable can be rationally defended as far as anything can be. Only the religious believer understands political values as less than fully valuable, and even she does not understand the appeal to lesser values as a grudging or unfortunate accommodation. Rather, the religious believer deploys a fully religious argument defending the removal of religious value from the political realm. Even here, political values have as much justification as they can have, given the commitments of the particular comprehensive doctrine.

The only sense in which Rawls's solution to the problem of stability might be understood as less than fully rational is that it will not be possible for us to understand every other citizen's commitment to reasonable political justification as itself fully rational. If overlapping consensus obtains, then each of us will understand her own commitment to reasonable political justification as fully rational; and we shall each be able to understand the commitments of others who affirm our own comprehensive doctrine in the same way. But there will always be other citizens, holding other comprehensive doctrines, whose accounts of their own political commitments will remain, for us, rationally opaque. We shall have no reason to doubt the reality or the sincerity of their political commitments, but we shall not be able to understand the ways in which they arrived at those commitments as themselves fully rational.

---

[18] "No one accepts the political conception driven by political compromise" (*ibid*, p. 171)

Is this a problem? It would be a problem only if the ends of political life or of political philosophy required that we converge on a fully comprehensive conception of rationality. But neither of these is the case. Political life requires that we have a shared framework to regulate our social life and to justify the use of coercive authority. Political philosophy, in turn, requires that we justify this shared framework in rational terms. If overlapping consensus obtains, each of these conditions is satisfied. All of us allow the idea of the reasonable to ground our political claims, and each of us regards this grounding as rationally sufficient. As for our fellow citizens, all we can ask is that they freely and willingly comply with the political claims that reason has justified. If overlapping consensus obtains, this is indeed the case. To ask for more than this, to ask that all others justify their free and willing commitments with the same reasoning that we do, is to make a confused demand, one that has no place either in politics or in philosophy. In politics, the demand for reasons is out of place; what we demand, even if we have reasons, is common behavior. In philosophy, the demand for identity is out of place; what we demand is that everyone give reasons, not that everyone act or think as we already do. To insist that everyone justify their free and willing political commitments in the same way that we do is to insist that reasoning must conform to a unique and preordained pattern; this is neither politics nor philosophy, but simply fanaticism.

IV. CONCLUSION

We are now in a position to provide a decisive answer to Habermas's objection. The objection was that Rawls's political liberalism lacks the force of a fully justified normative theory, since it appeals only to such pragmatic considerations as consensus and stability. On this view, Rawls is concerned more with securing political agreement than with justifying the terms of the political agreement we ought to secure. He makes what is at best an extrinsic concern of normative theories central to normative theorizing itself.

This reading assumes that Rawls's concerns are fundamentally instrumental: public disagreement is a problem in need of a solution, and political liberalism is able to provide that solution. But this is not the way Rawls proceeds. His concerns are precisely those of a traditional normative theorist: to spell out the necessary and sufficient conditions for political justification. The fact of reasonable pluralism is not itself a problem in need of a solution, but the natural consequence of rational activity under conditions of modern pluralism. If we then want to see our normative theory as itself the product of reason, if we want to avoid a conflict within reason, we must recognize reasonableness as a necessary condition of political justification. Under conditions of modern pluralism, Rawls is saying, no political theory is rational unless it could be accepted by anyone, no matter what comprehensive doctrine she affirms.

To avoid a detailed, exhausting, and potentially inconclusive survey of particular comprehensive doctrines, Rawls then proposes that we understand the reasonable as not just the necessary but also the sufficient condition for political justification. This is the only pragmatic step in the argument, the only time that Rawls appeals to instrumental considerations to justify the structure of his theory. But it is crucial to see that even here—especially here—Rawls does not regard the pragmatic appeal as providing a full justification for using the reasonable as the sole standard of political justification. The pragmatic appeal motivates this use, but precisely because the reasonable is not yet a sufficient condition of political justification, the pragmatic appeal introduces the problem of stability as a fundamental concern of the theory. To solve the problem of stability, particular comprehensive doctrines must deploy their own full conceptions of rationality. They must show, individually, that the reasonable is rational as a standard of political justification. If they can do this, then the appeal to reasonableness is not just a practical device to achieve agreement, but a sufficient condition of political justification.

In this sense, Habermas's objection gets Rawls exactly wrong. For Habermas, the reasonable is fundamentally a practical device to achieve agreement, and the rest of the theory is built around this idea. But for Rawls, the status of the reasonable as a practical device is only an intermediate and necessarily transitory stage in the development of the theory. This is why it is crucial to see the ideas of consensus and stability as deployed in a serial fashion. If we assume that these ideas are deployed together, we shall see them both as referring to the single and (supposedly) fundamental difficulty of securing agreement. But, in fact, they refer to the separate and intrinsically rational considerations that surround the practical problem of securing agreement. Consensus is important for the theory because availability to consensus is a necessary condition for political justification under conditions of reasonable pluralism. It is on rational and not on instrumental grounds that Rawls claims that modern political justification must be reasonable. The practical difficulty of achieving this is what prompts him to propose that reasonable justification is all the political justification we need. But this proposal is not justified by its practical success in securing agreement. It is motivated by its potential for practical success, but it is justified by its ability to serve as the object of an overlapping consensus of comprehensive doctrines, each of which regard it as a rationally sufficient basis for political justification. Stability is thus important for the theory because we still need to show that reasonableness is a sufficient condition of political justification.

Hence it is simply false to say that Rawls's fundamental concern is with securing agreement. The problems of consensus and stability refer not to the difficulty of securing agreement, but to the tasks of determining the necessary and then the sufficient conditions for political justification. If we can show that reasonable political justification is both rationally necessary and rationally sufficient, what more could we ask from our normative political theory?

This is not to say that there are not controversial and perhaps even unfounded premises in Rawls's argument. The claim that comprehensive doctrines are themselves exercises of reason is a very strong one; this is a thought that is extraordinarily difficult to maintain when reflecting on religious traditions other than one's own. The content of a theory of justice founded entirely on the notion of the reasonable remains very much in doubt. Finally, the appeal to an overlapping consensus is only a historical hypothesis, one that the theory is powerless to confirm. It may be that both religious and secular liberal doctrines are finally unable to converge on the notion of the reasonable. But the fragility of these premises provides no grounds for claiming that Rawls has abandoned the project of normative theorizing. On the reading I have given, if the premises are true, then Rawls has constructed an argument in which the notions of consensus and stability have all the normativity they need.

LARRY KRASNOFF

Reed College

PROPERTY COUNTERPARTS IN ERSATZ WORLDS*

Contemporary philosophers make free use of the concept of a possible world. Unfortunately, no satisfactory theory of possible worlds has been developed. David Lewis[1] has produced the most complete and convincing criticisms of the extant theories, along with his own startling account of the nature of possible worlds. Lewis thinks that I and my surroundings are part of the actual world. I can knock on bits of the world and hear the resulting sound. Anything that is spatiotemporally connected to a part of this world is itself a part of this world. Moreover, other worlds are of the same kind as this one—they are made up of concrete things, and all and only those things which are spatiotemporally connected to a part of a world is a worldmate of that part. I shall not be attacking Lewis's positive account,[2] but rather answering Lewis's criticisms of the view that he takes to be the best competitor to his own, thereby producing an acceptable theory of possible worlds. He calls the theory ***linguistic ersatzism***, presenting it as carefully and sympathetically as he can in or-

---

* This paper grew out of a series of discussions with Michael Thau many years ago. Although I must take responsibility for all that is wrong with this paper, he may well be responsible for much that is right about it. My thanks also to David Lewis, my colleagues at Southern Methodist University, and to the audience at the IUC Conference on Metaphysics in Slovenia, especially Christopher Gauker, Nenad Miscevic, Howard Robinson, David Sanford, and Johanna Seibt. I am most grateful to Doug Ehring, who read drafts of various parts of this paper. Finally, I am grateful to the National Endowment for the Humanities for funding this project

[1] *On the Plurality of Worlds* (Cambridge: Blackwell, 1986). Peter van Inwagen has presented an ingenious defense of what Lewis calls ***magical ersatzism*** in "Two Concepts of Possible Worlds," *Midwest Studies in Philosophy*, XI (1986): 185-213.

[2] Here are three forceful attacks on Lewis and a few authors who have proposed them: John Bigelow and Robert Pargetter argued persuasively that it should be possible for there to be two universes that are not spatiotemporally related to one another—"Beyond the Blank Stare," *Theoria*, LIII (1987): 97-114. Lewis must count these two universes as two worlds, and, hence, must deny that there is a single possibility that contains both. Lewis suggests that perhaps the unrelated universes are not possible and we only think they are because we are thinking of some other possibility (*op. cit.*, pp. 71-72). In a related criticism, Michael Jubien has argued against Lewis's account on the grounds that, if there were all the concrete universes Lewis takes there to be, that would still be a contingent matter, whereas Lewis must take it to be necessary—"Problems with Possible Worlds," in David F. Austin, ed., *Philosophical Analysis* (Dordrecht: Kluwer, 1988), pp. 299-322, especially p. 305 Perhaps the most common basis for rejecting Lewis's theory is that it just sounds plain crazy. Pargetter and Bigelow defend Lewis against this attack, but their defense depends on the assumption that the "blank stare" is all there is to the criticism. Perhaps the stare is a manifestation of the critic's sense that there is some deeper criticism that she is unable to verbalize

0022-362X/98/9506/293-316

der to show that even then it fails. I shall help myself to most of his suggestions on behalf of the linguistic ersatzist and argue that some further modification provides a version that can withstand the criticisms Lewis raises. The result will be at least the beginnings of a plausible theory of worlds.[3]

There is no particularly good terminology to use in the debate between Lewis and his opponents. The dispute is often described by saying that Lewis believes that possible worlds are *concrete* and his opponents that they are *abstract*, but Lewis points out that the concrete/abstract distinction is less clear than we would hope. He poses it as a dispute between *realists*, like himself, and *ersatzists*. But many of the latter would agree that there really are possible worlds, so it is hard to see why they are not realists about possible worlds, too. We might join Robert Stalnaker[4] in saying that Lewis is a more extreme realist (*ibid.*, p. 44) but, as Lewis points out, it is hard to find a dimension along which to measure in our attribution of extremism (*op. cit.*, p. 2, note). I shall continue, hesitantly, to use the concrete/abstract terminology and, equally hesitantly, the realist/ersatzist terminology.

Deeper, more obfuscating terminological difficulties remain. What is at issue are three distinct kinds of things, or purported kinds of things, each having a reasonable claim to the title *world*. Lewis's worlds are composed of concrete things (as well as whatever abstract things there may be). Everyone involved in the debate agrees that there is at least one of these things. Ersatzists say there is only one. Let us call that one thing the *concrete world*. Lewis would call it the *actual world* and say that there are many others that are nonactual. If there are several universes that are spatiotemporally unrelated to one another, as Lewis says, ersatzists will say that they are all parts of this one concrete world.

There are many different ways that this concrete world could have been and one way that it is. If we take the previous sentence at face value, and take its ontological commitments seriously, we end up taking *ways the concrete world can be* to be entities in their own right. Ways things are would seem to be properties of the things. The way the entire concrete world is would be a maximal-consistent property of the concrete world, and ways it could be would seem to be uninstantiated maximal-consistent properties. I shall call these purported entities *property worlds*. If there are such things, they are abstract entities. Lewis would prefer to take other concrete worlds to be the ways that this concrete world could be, and this concrete world itself to be the way it is. He does not think that there are uninstantiated properties—or at least none of the sort that would be required in order to replace his concrete possible worlds[5]—and even among ersatzists there will be debate over whether there are uninstantiated properties of the right sort.

---

[3] Just as my account is a development of Lewis's suggestions, his proposals on behalf of linguistic ersatzism are themselves developed from Willard Van Orman Quine's "Propositional Objects," in his *Ontological Relativity and Other Essays* (New York: Columbia, 1969), pp. 139–60. The theory I end up with is also similar to David Armstrong's *A Combinatorial Theory of Possibility* (New York: Cambridge, 1989), though Armstrong has no acceptable answers to Lewis's criticisms. He ends up forced to bite bullets that need not be bitten. In a recent conversation with Ted Sider, I learned that he is working on a theory very similar to mine, but the two theories were developed entirely independently of one another.

[4] *Inquiry* (Cambridge: MIT, 1987).

There are also descriptions of this concrete world. The linguistic ersatzist identifies worlds with maximal-consistent descriptions of the concrete world, one of which is true and the rest false. These descriptions need not be in English, or any other natural language. For instance, Lewis suggests that the linguistic ersatzist could use mathematical representations—ultimately, sets. Although these entities may be somewhat controversial, they are much less controversial than either uninstantiated property worlds or the numerous concrete worlds Lewis posits. The linguistic ersatzist's job is to show that these mathematical entities can do the work that worlds are meant to do. I shall call these things *representation worlds*. Lewis argues that representation worlds cannot do the job, and my task will be to undermine his criticisms. I shall not get involved in a debate about whether representation worlds should be called *worlds*. I shall typically use the word 'world' without any attached adjective if the context makes it clear to which of the three kinds of entities I am referring; and I shall henceforth use 'ersatzism' to refer to linguistic ersatzism.

To get ersatzism off the ground, Lewis starts with the Quinean characterization (*op. cit.*, p. 147).[6] A representation world is a set that can be used to assign, in the simplest version of the theory, fullness or emptiness to each point in a space-time manifold. Such a world, then, is a set of sets. Each of these member sets is itself an ordered pair,[7] having as a first member an ordered quadruple of numbers—representing a particular point in the space-time manifold—and having as its second member either a 1 or a 0—representing fullness or emptiness. Without entering into the debate about whether numbers are sets, we can at least note that they can be replaced by sets. So this account is committed to nothing any more ontologically suspect than sets.[8] A visual representation of a Quinean world looks something like this:

---

[5] Lewis allows that, with our quantifiers wide open, there are properties that are instantiated at other worlds but not at this one, but this depends on the other worlds being the sorts of things he takes them to be, and not the sorts of things the ersatzist takes them to be

[6] Lewis cites Quine's "Propositional Objects," making it clear that neither Lewis nor Quine adopts the "Quinean" characterization.

{<<13, 69, 1560, 2>, 1>, <<589, 4, 37, 38>, 0>, <<77, 82, 99999, 311>,1>, ...}

The account must be more complicated in order to allow for more properties and for differently structured manifolds. So a world is still a set of sets, but each member ordered pair has as its first member an ordered *n*-tuple of numbers representing a point in the manifold and has as its second member a set of numbers representing properties.[9] A visual representation of a modified Quinean world (this one with a two-dimensional manifold) looks something like this:

{<<13, 69>, {3, 48}>, <<589, 38>, {7, 19}>, <<88, 235>, {3}>...}

Of course, these sets do not represent intrinsically, but only relative to an interpretation. This poses no difficulty as long as the rules of interpretation can be specified in a finite vocabulary, which they can be. Lewis objects neither to the existence of these sets, nor to their need for an interpretation. He poses two objections to this ersatzism, each of which takes two forms (*op. cit.*, pp. 150-65).

Lewis's first objection is that this account of possible worlds cannot be used to reduce modality to something nonmodal. In order to complete the account adequately, some modality will have to be assumed. This shows up first in the need to select from among all of the sets that are of the right structure to be representation worlds those sets which, under the appropriate interpretation, are maximal and consistent. The modal notion of maximality can be defined in terms of consistency, but that still leaves us with consistency as a modal primitive. The second point at which primitive modality seems required is when the ersatzist tries to give an account of implicit representation. The more sparse our choice of properties to be assigned locations in each world, the less we can say explicitly about the world. For example, if we explicitly say where the properties of subatomic particles are instantiated but do not explicitly say where donkeyhood is instantiated, we must get truths about donkeys by way of their being implied by the truths about the properties of subatomic particles. This sort of implication seems modal, requiring that the truths about donkeys are necessitated by the truths about particles. So we seem stuck with implication as a modal primitive as well.

---

[7] I assume here an account of ordered *n*-tuples that makes them reducible to unordered sets.

[8] Because a different coordinate system can be used to describe the same distribution of matter, what we really need is not the sets just described, but equivalence classes of such sets. This same point applies to the modified Quineanism I discuss below. I shall ignore this complication wherever possible.

[9] There is another complication that I shall not pursue. The ersatzist needs some way to represent the existence of abstract entities that are not located. This is one of the reasons that this paper is only the beginning of a plausible theory of worlds.

Lewis's second objection is that the linguistic ersatzist is unable to distinguish all of the possibilities that need distinguishing. The first possibilities that get omitted are those involving a single individual's possibility of being a specific one of several duplicates. The ersatzist can only distinguish one possibility from others by their properties. This leaves her unable to account for possibilities indistinguishable from one another by their properties, and that is what the duplicates would be—indistinguishable possibilities for the one actual world individual. The second class of omitted possibilities are those which involve alien properties—properties that do not exist in the actual world. To say that a property could have been instantiated, the ersatzist needs a way to represent the instantiation of that property. To represent that, she needs some way to refer to that property. Moreover, the reference is actual reference, so it must be accomplished with tools that are actually available. This is a difficult task when the property being referred to does not exist in the actual world.

I shall attempt to answer only one half of each of these two objections. With respect to the problem of primitive modality, I shall not discuss the problem of implicit representation, because I have done so elsewhere.[10] The half I shall deal with is the problem of consistency. Most of this problem can be shifted over to that of implicit representation, if we are miserly enough about which properties we represent explicitly; but there will still be an ineliminable remainder to the consistency problem concerning those properties which are determinates of a single determinable. It is that remainder that I shall discuss here. I shall also discuss only half of the problem of undistinguished possibilities. The half I shall not discuss is the duplicates problem, because I do not feel that I have anything useful to say about it.[11] The half I shall discuss, and shall spend the bulk of the paper on, will be the alien-properties problem. It is to that problem that I now turn.

---

[10] "Ersatz Worlds and Ontological Disagreement," *Acta Analytica*, XV (1996): 35-44. The present paper, the one cited here, and "Five Layers of Interpretation for Possible Worlds," *Philosophical Studies* (forthcoming), all evolved from a single paper that was presented at the IUC Conference on Metaphysics in Slovenia in June 1995. *Acta Analytica* is the journal representing that conference.

One's approach to alien properties will depend upon one's view of the nature of properties. For instance, as Lewis notes (*op. cit.*, p. 160), if we accept that properties are abstract entities in a Platonic heaven, existing independently of being instantiated, then we can say that any property that could exist actually does exist. Now, we can take representation worlds to have these abstract entities as ingredients, since we are allowed to use anything there is as an ingredient. Where our modified Quineanism took a representation world to be a set of numbers only, we can instead replace those numbers which represented properties with the properties themselves. A world—still a set of sets—that includes as one of its members a set with a location L as its first member and a set containing property P as its second member represents P as being instantiated at L. The possibility of some other property being where P is gets represented by replacing that set with one that includes that other property and L. The problem of alien properties simply goes away, because there are no alien properties.[12]

But Lewis does not accept this view of properties. He thinks that the ersatzist will be unable to make good sense of uninstantiated properties (*op. cit.*, pp. 160-61).[13] So properties that are not instantiated in this world do not exist in this world and are therefore unavailable to use as ingredients in other worlds. An ersatzist who does not embrace uninstantiated properties must look elsewhere for a solution to the problem of alien properties. I find Lewis's dismissal of uninstantiated properties plausible, so I shall not fight him on this score.

---

[11] I thank Lewis for steering me away from some misguided things that I thought of saying about this problem. He suggested that the ersatzist's best response to this problem is just to deny that a world of infinitely many Napoleon duplicates, for instance, does provide infinitely many possibilities for Napoleon. I am prepared to adopt this suggestion for the time being, though Lewis thinks it a high price to pay. (This is the only portion of my paper that I have discussed with Lewis. No doubt he would have found difficulties in other parts if he had seen them.)

[12] This still counts as a form of *linguistic* ersatzism for Lewis, because we can take the properties to be names for themselves. But if we adopt this view of properties, we could, it seems, opt for property worlds instead of representation worlds—a possible world just is a property of the one concrete world, and all but one of the possible worlds is an uninstantiated property.

[13] Lewis thereby agrees with Armstrong, *Nominalism and Realism* (New York: Cambridge, 1978).

Lewis suggests the ersatzist try Ramsifying: instead of using names of properties in our representation worlds, we can existentially generalize over the properties (*op. cit.*, p. 162). He does not go into ontological detail here, but an existential generalization can itself be represented by a set of sets of numbers. Ontologically, representation worlds are the same sets that I proposed above. For example:

$$\{<<13, 69>, \{3, 48\}>, <<589, 38>, \{7, 19\}>, <<88, 235>, \{3\}>\ldots\}$$

But now those sets are given a different interpretation. The numbers in the property places should now be interpreted as existentially bound variables instead of names. This is an improvement over my earlier description anyway, since I gave no principles for attaching a number to a particular property for which it was supposed to be the name. The interpretation should stipulate that when two or more member sets of a single world contain the same number in their property places, that number represents the same property, and different numbers represent different properties.[14] On this Ramsification approach, a world provides a complete description of the distribution of properties—it tells us, for each property, where it is instantiated. Lewis's version of the Ramsification proposal has the ersatzist explicitly describing the nomological role of each property, but if we adopt the Humeanism of which Lewis is fond, nomological role will be derivative upon the distribution of properties. Although a world completely describes a distribution, it says nothing about which properties are being distributed. That is, it does not tell us anything about the transworld identification of properties.

There is no longer a problem of alien properties, because that problem only arose when we were using names of actual properties in the construction of the worlds. On the original proposal, we had no way to

---

[14] So the set in the text, under its new interpretation, is equivalent to something like the following in a more familiar notation (taking '*Lx*' as '*x* is a location', '*AFx*' as '*F* is at *x*', and pairs of numbers as names for locations):

$$(\exists F_3)(\exists F_7)(\exists F_{19})(\exists F_{48})\ldots(L<13,69>\& L<589,38>\& L<88,235>\ldots\& AF_3<13,69>\& AF_{48}<13,69>\& AF_7<589,38>\& AF_{19}<589,38>\& AF_3<88,235>\& <13,69>\neq<589,38>\neq<88,235>\ldots\& F_3\neq F_7\neq F_{19}\neq F_{48}\ldots)$$

(Part of the reason for my 'something like' qualification is that, in treating the pairs of numbers as simple names, we lose some of the information about the locations which is conveyed by treating them as coordinates.) A world in which there is nothing more filling in the ellipses is stipulated to be complete. In the second-order predicate calculus a clause would be added to the effect that:

$$\& (x)(F)((x=<13,69>\vee x=<589,38>\vee x=<88,235>)\&(F=F_3\vee F=F_7\vee F=F_{19}\vee F=F_{48}))$$

I give these translations for the sake of clarification, but I emphasize that the "sentences" receiving interpretations according to my ersatzists are composed of sets, not inscriptions.

name an alien property, and, hence, no way to represent its instantiation in a given world. What makes Ramsification useful is that it allows us to pick out properties in other worlds by description, rather than by name. Adopting Ramsification, we can now say that there is a property above and beyond any of the actual properties, and that it is distributed thus and so, without having to say what property it is.

But new problems arise to take the place of the old ones (*op. cit.*, p. 162). One consequence of the Ramsification approach is that same distribution entails same world. Consider a very simple world *w* with just two properties, call them *dis* and *dat*. Now consider a world *w′* that is the exact inverse of the first—the properties have exchanged places. On the present proposal, it turns out that there is really just one world here. A Ramsey-style description of *w* will exactly match the description of *w′*; and for representation worlds, the description is all there is to the world. So, again, it seems that there are possibilities that get lost on the ersatzist program. Moreover, lest it be thought that some combination of property names and existential generalizations can avoid this problem, we need only consider a version of the inverse-worlds problem as applied to two or more alien properties—properties for which we have no names.

To resolve this problem, the ersatzist might insist that there are no inverse worlds (*op. cit.*, pp. 162-63). One way to insist upon this is by saying that properties are world bound—no property exists in more than one world. For my ersatzist, this is to say that no property is represented by more than one world. Instead of simply failing to say anything about transworld identity in our interpretation of the sets that are representation worlds (which would have left the worlds incomplete anyway), we stipulate that the properties in one Ramsey sentence are not identical to any in the other Ramsey sentences. There is no possibility of disness and datness switching places, because any world that we would be tempted to describe by saying that disness and datness had switched places does not really contain disness or datness at all.

But now, it seems, we have made too much essential to properties. It should be true that some particles would have been positively charged even if I had never written any sentence about positive charge. But on the present proposal, that seems to come out false because positive charge only exists in this world; any world in which I failed to write about positive charge would be one in which positive charge does not exist. Likewise, the whole point of the inversion objection is that it seems true that dis and dat could have been inverted, yet the present proposal seems simply to deny this. So the suggestion that properties are world bound does not sound like a response to the inversion objection, but only a rejection of the intuitions behind it.

Here, I suggest a modification to ersatzism that Lewis did not consider. Just as he is a counterpart theorist about individuals, the ersatzist can be a counterpart theorist about properties.[13] According to Lewis, no individual exists in more than one world. Yet this does not make all of the individual's properties essential to it. De re modal discourse is to be understood, not in terms of the actual individual's properties in other worlds, but in terms of its counterparts' properties in other worlds. Marlon could have been a contender, if and only if Marlon has a counterpart in some world who is a contender. Likewise, I propose, even if we insist that no property exists in more than one world, we do not have to say that all of a property's properties are essential to it. Positive charge could have existed even if I had never written about it, if and only if positive charge has a counterpart that exists in a (close enough) world in which I (or rather my counterpart) do not write about that counterpart. Lewis's counterpart relation is grounded in similarity. Marlon has a particular individual in another world as a counterpart, if and only if that individual is similar enough to Marlon and more similar to Marlon than anything else in that world. This similarity relationship is nontransitive and context sensitive, and this provides much of the strength of counterpart theory. Likewise, I propose that the counterpart relation for properties be grounded in a nontransitive, context-sensitive similarity relation.

It is not easy to see how to compare properties. Is positive charge more similar to spin up than it is to negative charge? When I start thinking about this question seriously, I start to think about the roles of the various properties in question. So I propose that the similarity relation between properties be understood as a similarity relation between their roles. To describe a property *P*'s role completely, we say 'it is such that...', where the ellipsis is filled in with the rest of the description of the entire world: *P* is such that it has such-and-such a distribution among other properties *P*1, *P*2, and so on, that have so-and-so distributions. Where a world is a Ramsey sentence (a set of sets interpreted as an existential statement), a property's role in that world would be the open sentence that results from dropping the existential quantifier that binds that property. Given our assumed Humeanism, such a characterization of a property's role also characterizes its nomological role. In the end, to compare property roles (and thereby to compare properties) is to compare entire worlds; and let us not forget that this means that a comparison of property roles is ultimately a comparison of sets.

---

[13] I have also declared myself a proponent of counterpart theory for individuals in, for instance, *The Ontology of Physical Objects* (New York: Cambridge, 1990); see also my "Varieties of Four-Dimensionalism," *Australasian Journal of Philosophy*, LXXI (1993): 47-59.

The easiest way to compare sets is to consider which members they have in common. Consider two worlds that are exactly alike, have exactly the same sets as members, with the single exception that where one of them has a property number nineteen paired with a certain location, the other has a twenty-five paired with that location. Clearly, these two sets, considered as sets, would count as very similar in many contexts. Likewise, considered at a higher level of interpretation, considered as representations of property distributions, these two representations would often count as very similar. Independent of any question of which properties are being distributed, the *patterns* of property distribution are very similar, and there are surely many contexts in which we would therefore count the two worlds as very similar. In such contexts, we would say that the property numbered seven in the one world plays almost exactly the same role as the property numbered seven in the other world, and we would even say that the property numbered nineteen in the one world plays almost exactly the same role in the other world. In contexts in which these are the correct similarity judgments, property seven in the one world is the counterpart of property seven in the other, and property nineteen in the one is the counterpart of property nineteen in the other. Worlds similar to the actual world in the same way (though not to the same extent) as these two worlds are to each other ground judgments like 'positive charge could have existed without my writing about it' and even 'there could have been a single isolated incident in which two positively charged particles attracted one another'.

Another way that properties in different worlds can be counterparts of one another is by playing similar nomological roles. The laws of nature in a world are, according to our assumed Humeanism, a function of the distribution of properties in that world. Worlds with the actual world's laws may have radically different distributions. In some of these latter worlds, there will be properties that play the same nomological role as the actual world's properties. They will be counterparts of the actual properties. Worlds like these also ground our judgments like 'positive charge could have existed without my writing about it'. On the other hand, worlds that are nearly identical in distribution are likely to have different laws. For instance, distributions so similar as to ground our judgment 'there could have been an isolated incident in which two positively charged particles attracted one another' will have to have radically different laws of nature to account for the isolated difference of property distribution. Because there are contexts in which we may recognize this and still count the properties as very similar, the two described worlds can ground judgments like 'positive charge could have existed even if the laws of nature were radically different'.

But because there are other contexts in which laws of nature are important to us, we sometimes make judgments like 'there could not have been even a single isolated incident in which two positively charged particles attracted one another'. That is, in these contexts, those very same worlds do not contain counterparts of positive charge precisely because that would require such a large difference in laws. The fact that we are sometimes led to pass the judgment that there could be isolated incidents of positive attracting positive and sometimes to pass the opposite judgment is some evidence for a context-sensitive counterpart account of properties. One more judgment that might be supported by the same pair of worlds is that miracles are possible (by which I mean that the laws of nature of a given world could fail to be exceptionless at that world). Identifying the properties of the one world with those of the other—and noting that in all regions other than the one in which the isolated incident occurs, the distributions are identical—we might identify the laws of the two worlds. Then the one isolated incident would be counter to the laws of its own world, thus grounding the judgment that laws need not be universal.

There are also worlds that are radically dissimilar to each other using any natural or intuitive similarity weighting. When we compare the patterns of property distribution, we see no similarity; and when we compare the laws that would be at work in the two worlds, we see no similarity. Depending on the rest of the details, we may have no temptation to count these worlds as containing property counterparts of one another. This grounds something that seems right to me: while we do not want to count all of a property's properties as essential to it, we also do not want to say that just anything goes. We do not want to say that all of a property's properties are accidental to it.

Where do we draw the line between which properties are accidental and which are not—the line between which other-worldly properties are similar enough to count as counterparts of an actual property and which are not? I think that is a vague, context-relative matter. For instance, we may initially consider two radically different worlds, *A* and *Z*, to have no property counterparts. But suppose we compare them again as two members of a series of worlds with *A* as the first member of the series, *Z* as the last, and all the worlds in between ordered so that each world is very similar to its neighbor. Then we may be drawn to saying that these worlds do have property counterparts after all (not because we are accepting transitivity, but because our mind is drawn along from *A* to *Z*). This also lends some support to a counterpart account of the transworld relationships between properties. If a property's essential properties were grounded in transworld identity, there would be no room for this sort of context relativity, since identity is not relative.

In normal contexts, I am not inclined to say that positive charge could have played just any old role. In particular, it seems to me that one role it could not play is the one that is actually played by negative charge. Thus, I do not think that there is the possibility of positive and negative charge switching roles, and I do not have to produce a world that represents such an inversion. I am prepared to admit, however, that for much simpler properties, properties that play much simpler roles, properties that exist in much simpler worlds, perhaps role switching is possible. The simple properties that I earlier labeled dis and dat would be examples of this if any were.

If I accept that 'dis' and 'dat' in world *w* could have switched places, I must have some way to represent their doing so. I do not need to posit an inverse world—a world in which dis and dat are each playing the other's *w* role—because I am a counterpart theorist about properties. What I need is a world *w** in which a counterpart of dis plays dat's *w* role and a counterpart of dat plays dis's *w* role; and there is such a world; it is *w* itself. World *w* contains two properties, the roles of which are similar enough so that each can be the counterpart of the other. If one insists, contrary to my own intuitions, that positive and negative charge could be inverted, I shall then say the same for positive and negative charge. The actual world itself is the world that grounds the possibility of inversion. If the roles of positive and negative charge are similar enough to make inversion possible, they are similar enough to allow each property to be the counterpart of the other.

So the ersatzist can solve the alien-property problem. She should Ramsify properties while holding that no property exists in more than one world, and then she should adopt a counterpart theory for properties in order to allow us to say true things about what is possible for properties. We can think of representation worlds as each telling a story. They tell stories of the distribution of properties. All but one of these stories is false. The stories use fictional names—they are not the names of any real objects (in effect, making them variables). We can think of counterpart theory in terms of readings of the various stories.

When we read a story, we might "recognize" one of the characters as a real individual, even though some of the things that are said about that character are false. But if too many things said about her are false, then we do not identify her as the real individual in question. Likewise, when we read a story about properties, we can identify those properties with real properties—recognize them—even though they are not exactly like the real properties. But if the properties in the story are too radically different from the real ones, then we will not count the fictional properties as being the real ones. If we want, we can even *talk* of transworld identity. An object or property in one story—one world—*counts* as the same object or property from another story, if and only if it is similar enough and there are no others that are more similar. The *counting as the same* relation is just the counterpart relation in another guise, and it is nontransitive and context relative. Notice that comparing worlds is like comparing written stories in this respect: we can compare written texts by comparing the actual distribution of ink on the pages, but we often compare them at a higher level of interpretation.[16]

We can think of counterpart theory as adding yet another layer of interpretation to our possible worlds. So far we have been interpreting sets of the right structure as being Ramsey sentences describing property distributions. I suggested that we can then identify some properties as counterparts of properties at other worlds. We can count properties in one world as identical to properties in another world. This can be thought of as a higher level of interpretation. The same world that I had previously said does not represent positive charge, I can now say, at this new level of interpretation, does represent positive charge. It accomplishes this representation not in virtue of a stipulated assignment of that property to one of the property numbers in the set, but in virtue of our similarity judgments. Likewise, given the property identification, we can also identify individuals across worlds at a higher level of interpretation. We can interpret certain distributions of certain identified properties as being individuals, and particular individuals that exist in other worlds. This does not mean that there really is transworld identity. The *counting as identical* relation cannot be identity, since it lacks the formal properties of identity. But at this higher level of interpretation, in a given context, a world can represent actual individuals and properties.[17]

---

[16] This talk of stories may mislead the reader into mistakenly associating my theory with Gideon Rosen's "Modal Fictionalism," *Mind*, xcix (1990): 327-54. My worlds are fictions about the one concrete world (except for the actual world, which is non-fiction). The fiction involved in Rosen's account is a fiction about possible worlds. His fiction is Lewis's theory of possible worlds—when asserting that *p* is possible we pretend to accept Lewis's false theory of possible worlds and to assert that *p* is true at one of those worlds. I say there really are possible worlds, and they are *fictions*. Rosen says that there are no possible worlds; they are merely *fictional*, themselves being characters in Lewis's fiction

Although I have spoken of comparing properties and property roles across worlds, it must be emphasized that ultimately all of these comparisons are between items that exist in the actual world; they are not comparisons between mere possibilia. We compare one representation with another to see how similar they are, and these representations actually exist. Just as we can compare two characters or properties in two stories without having to suppose there really are any such things as the characters or properties in question, we can compare individuals or properties in different worlds without supposing that there really are such individuals or properties. In the one case, what we are really comparing are stories; and in the second, what we are really comparing are worlds. In the first case, the stories are real, actually existing objects, even if the characters depicted are not; and in the second, the worlds—the sets—are real, actually existing objects, even though the characters they depict are not.[18] Because the similarity comparison is between actually existing objects, there is no need to suppose that the respects in which the various individuals or properties are similar are themselves properties that exist in more than one world.[19]

---

[17] My "Five Layers of Interpretation for Possible Worlds" focuses more directly on the various layers of interpretation that are involved when worlds represent. We can, if we want, add one more layer of interpretation, by moving the 'counts as the same' relation into the interpretation itself. We can then say that a *fully* interpreted world really does represent actual world objects and properties. In this case, there is transworld identity across fully interpreted worlds. It is the same object that exists according to the different stories/worlds. Transitivity will be preserved across fully interpreted worlds (though it may not seem to be if conversational pressures yield shifting interpretations), and the identity relation will not be context relative across such worlds, because the context would have completed its job by the time the worlds are fully interpreted. This is the line I take in my "Linguistic Ersatzism, Counterpart Theory, and Trans-world Identity," an American Philosophical Association presentation (Central Division, April 1997).

[18] Of course, we must be careful to distinguish the two senses in which a world can be actual. In the sense of being an actually existing object, all representation worlds are actual. In the sense of accurately representing the one concrete world, only one representation world is actual.

[19] This is why there is no problem of transworld identity for roles. My thanks to Nenad Miscevic for making me see that there is a problem to avoid here.

The fact that the similarity involved here is ultimately a relationship between representations makes for an important difference between my counterpart theory and Lewis's. For Lewis, a crossworld comparison is no different in principle from a crosstown comparison. This would make a counterpart theory for properties very counterintuitive. We are already used to the idea of exactly similar but distinct individuals, but not to exactly similar but distinct properties. In contrast, the ersatzist does not have us comparing individuals or properties at all, but only representations of them. We compare one representation with another to determine whether the stories are similar enough to count as representing the same individual or property. For Lewis, there is a real question as to whether his worlds are composed of the same items of which the actual world is composed. For the ersatzist, this question is wildly misguided. The worlds are composed of sets, not of rocks, people, or properties. The appropriate question will never be whether another world literally contains the same objects or properties as this world, but only whether it represents the same objects or properties; and it should not be at all surprising if questions about what a given representation represents are answerable only relative to context. Accepting that worlds are just representations makes it easier to accept counterpart theory for properties, and it may make it easier for some to accept counterpart theory for individuals as well.[20]

Before leaving behind the problem of alien properties and the counterpart solution to it, I should note that Lewis's explicit objection to world-bound properties is not that such properties would have the wrong essential properties. His objection is put directly in terms of his recombination principle (*op. cit.*, p. 163). One formulation of this principle is that "patching together parts of different possible worlds yields another possible world" (*op. cit.*, p. 87), and another formulation is that "anything can coexist with anything else, at least provided they occupy distinct spatiotemporal positions" (*op. cit.*, p. 88). In the end, he puts the principle in terms of duplicates of individuals, because that is the best way for him to make it fit with his counterpart theory (*op. cit.*, pp. 88-89). Because duplicates are explained in terms of property identity across worlds, I must propose a different version of the recombination principle, if I am to make it consistent with the counterpart theory of properties. I could use Lewis's with a new account of duplicates in terms of the shared subsets between two worlds. But I think the principle is better stated directly in terms of the distributions of properties:

---

[20] Douglas Ehring is the one who first suggested to me that property-counterpart theory is more plausible for an ersatzist than for a realist.

> For any number of dimensions, any number of properties, and any combination of the two, there is a world that represents that combination.

In short, any set with the modified Quinean structure is a possible world.

I turn now to the problem of primitive modality. Lewis considers two ways in which this problem arises, noting that the two can be played off against one another (*op. cit.*, p. 152). There is ineliminable modality in our account of possible worlds, if we need it in order to explain consistency (to distinguish a possible world from a set that would represent something impossible); and there is ineliminable modality, if we need it to explain implicit representation. Lewis points out that we can lessen the need for some of the primitive modality of the one kind by increasing the need for the other kind. For instance, if we have both bachelorhood and marriedness represented in the world-making language, then we need something to block a representation of their being coinstantiated. But if we explicitly represent only marriedness, then we do not run the risk of representing married bachelors. To go further, if we explicitly represent only simple, fundamental properties, then we do not run the risk of inconsistency unless the fundamental properties themselves can be inconsistent with one another. For the most part, this eliminates the need to make consistency primitive, but now we have no way to say explicitly that bachelorhood is instantiated within a given region. We must represent this implicitly. Thus, we can shift the primitive modality burden away from consistency and onto implicit representation.

This still leaves us with the primitive modality that is involved in implicit representation. I shall here only sketch my solution to that problem. Elsewhere,[11] I suggest that implicit representation is ultimately a linguistic matter. Once all facts about fundamental properties are represented, facts about bachelors are already fully represented. They are not further facts above and beyond the facts about distributions of fundamental properties. To get them explicitly stated we would need to know the rules for using the word 'bachelor', but it is not the ersatzist's job to state those rules any more than it is her job to state rules of translation from English to

---

[11] "Linguistic Ersatzism and Ontological Disagreement."

French. This is as far as I shall take the discussion of this problem here; to go further would be redundant and is not necessary for our present purposes.

What I have just suggested in order to deal with the primitive modality objection is to shift the burden to the problem of implicit representation and then to answer that problem with a restriction on which kinds of properties are represented by the numbers that represent properties. We cannot shift all of the burden to implicit representation, however, if it turns out that there are fundamental properties that themselves conflict. If that is so, then we must find some way to avoid representing the possibility of the two fundamental properties being coinstantiated. If we block such representations simply by appealing to the fact that the properties are necessarily not coinstantiated, then we are appealing to a primitive modality. This is what remains of the consistency problem.

Lewis suggests that we might have such cases of conflict when the fundamental properties are two determinables of a single determinate. He gives the example of positive and negative charge (*op. cit.*, pp. 154-55). It is not nomologically possible for these two properties to be coinstantiated. Is it possible in any broader sense: Is there any world at all in which they are coinstantiated? Lewis does not take a stand on this question. His point is that the ersatzist is faced with the seemingly futile task of answering this question in order to complete her theory. She cannot merely plead ignorance. The problem is not that the ersatzist must in general know what is possible in order to give an account of what possible worlds are. The problem is that there are many sets that have the right structure to satisfy the ersatzist's description of possible worlds and that, if she must deny that they all really are possible worlds, then she must have some criteria for distinguishing some of these sets from the others. Without such criteria, her account of what a possible world is is incomplete. If we can represent the coinstantiation of these properties, then without such criteria she does not have an account of what distinguishes a possible world from other sets. If the only way to complete the account is to appeal to a modal notion, then she is stuck with primitive modality.

My response to this objection is built upon the counterpart theory for properties. If we think that what it means for two properties to be coinstantiable is for them to be coinstantiated in some world, then the ersatzist will have to decide whether a set representing positive and negative charge as instantiated in the same location is a world. But once counterpart theory is adopted, the ersatzist can say that, at the level of interpreting sets as Ramsey statements about property distributions, there is no representation of those very properties being coinstantiated, while also saying that, at the level of interpreting the distributed properties as counting as the same as properties in other worlds, they are coinstantiated. The ersatzist can deny real coinstantiation—coinstantiation involving real transworld identity—in any world without denying coinstantiability, because it is not transworld identity that matters for coinstantiability, but counterparthood. The question of coinstantiability turns into the question of whether there is any world in which a counterpart of positive is coinstantiated with a counterpart of negative; and that is a question about our similarity judgments.

There are many worlds in which many properties playing some of the roles of positive and negative are coinstantiated. The question is: How dissimilar can the roles get before we are no longer willing to count the coinstantiated properties as counterparts of positive and negative? One especially relevant question is whether the mere fact of coinstantiation is enough to rule out the counterpart relation in this case. The similarity judgments involved in answering these questions are themselves context relative. Indeed, it may even be that there is no answer to these questions except in a special context in which we are forced to judge; and that seems just as it should be for the issue of coinstantiability of positive and negative. The task of resolving the issue seems futile because we cannot even imagine the type of information that would decide the issue for us.

For instance, we cannot imagine what science could discover that would tell us whether the impossibility of coinstantiation is merely nomological or something stronger. That is a problem if we think that the answer to that question is a matter of science. But it is no longer a problem if we accept that the answer depends on our similarity judgments. Where we are unable to make such judgments, we can simply say that there is no answer. This seems to be added support for a counterpart theory for properties. Counterpart theory provides an explanation of the futility of the task, an explanation that is not available otherwise. The very fact that seemed at first to make life difficult for the ersatzist seems now to be evidence for my counterpart account of properties which is at the heart of my ersatzism.

So it turns out that absolutely any set with the right structure—sets of sets pairing property representations with location representations—is a possible world. There are no sets like these which represent the impossible. Such a set may seem to represent the impossible, if we assume that the property numbers in the sets are names for specific properties that exist in many worlds (the same property being named in several of the representation worlds). Then we would have the question of consistency, as in Lewis's positive/negative example. But if we accept counterpart theory for the fundamental properties, keeping each property world bound, then there is no risk of incompatible properties being represented as coinstantiated. Whatever properties are represented by a Ramsey sentence as coinstantiated are thereby compatible properties. The open question is not whether they are compatible, but whether they are enough like our properties of positive and negative charge to be counterparts of them.

We might put the point a different way, noting that one of the presumed virtues of counterpart theory is that it grounds anti-essentialism. A single object will have different essential properties in contexts in which it is thought of as a statue from those which it has in contexts in which it is thought of as a lump of clay.[22] Likewise, a virtue of property-counterpart theory is that it grounds anti-essentialism for properties. What is essential to positive charge will vary with which respects of similarity are important to us in different contexts. Just as there is nothing about the object itself that requires that it have its shape essentially, there is nothing about positive charge itself that should require that it has any specific relations to other properties essentially.

What is important to note is that this property anti-essentialism is not an antirealism about properties. The primary reason to believe in properties, to my mind, is to account for what is shared in cases of similarity. Properties are universals. Properties can do this job perfectly adequately even if what is essential to them is dependent on how we think of them. Indeed, property anti-essentialism is even consistent with David Armstrong's[23] scientific realism about universals. All Armstrong's realism requires is that the science of a given world tells us how to find the universals of that world; it does not require that the science of a given world tells us to which other-worldly universals the actual world's universals are identical. The fundamental properties of my worlds figure in the laws and participate in patterns of causal interaction (though, because of my assumed Humeanism, the laws and patterns of interaction will only appear at higher levels of interpretation), and so they should be discoverable in just the way that Armstrong would hope.

---

[22] The essentialist debate rages on. See, for instance, Quine, "Reference and Modality," *From a Logical Point of View* (Cambridge: Harvard, 1953), pp. 139–59; Saul Kripke, *Naming and Necessity* (Cambridge: Harvard, 1980), and Alan Sidelle, *Necessity, Essence, and Individuation* (Ithaca: Cornell, 1989).

[23] *A Theory of Universals* (New York: Cambridge, 1978).

Before closing with some possible challenges to my property-counterpart theory, I would like to acknowledge an earlier relative. The idea of analyzing modal claims about properties in terms of transworld similarity was briefly considered a few years ago by Denis Robinson.[24] In considering an argument by Lewis,[25] Robinson is led to consider an account of properties according to which they have their nomic roles essentially (call such an account *Shoemaker/Swoyerism*, or S/S[26]). It is in attempting to bring S/S into harmony with many of Lewis's commitments that Robinson suggests that property identity can be supplemented with property counterparthood. Robinson suggests that when intrinsic similarity is combined with sameness of nomic role, the combination is necessary and sufficient for property identity. Intrinsic similarity alone is sufficient for property counterparthood. Modal claims should now be understood in terms of both property identity and property counterparts. Thus, Robinson finds himself considering a theory that lies somewhere between S/S and my property-counterpart theory.

This intermediate theory is like S/S in accepting transworld property identity and tying it to nomic role. In contrast, my counterpart theory denies all transworld property identity (though recall footnote 17). What is new about the intermediate theory is that it allows that property counterparts have a role to play in interpreting modal discourse about properties. This is a move in the direction of property-counterpart theory, though it falls short of a full-fledged counterpart theory, because it is still committed to some transworld identity for properties and to granting such identity a role in interpreting modal discourse.

Robinson does not consider a full-fledged property-counterpart theory, and he does not endorse even the intermediate theory he does consider. He is pessimistic about the intermediate theory's prospects because of its use of intrinsic similarity to ground the counterpart relation: "a seemingly primitive unfathomable, ineffable, and arguably epiphenomenal kind of similarity" (*op. cit.*, p. 31). This worry does not apply to my counterpart theory because I ground the counterpart relation in property roles (which should not be confused with the nomic roles of S/S).

---

[24] "Epiphenomenalism, Laws, and Properties," *Philosophical Studies*, LXIX (1993): 1-31.

[25] The Lewis article that prompts Robinson's discussion is "What Experience Teaches," in William G. Lycan, ed., *Mind and Cognition: A Reader* (Cambridge: Blackwell, 1990), pp. 499–519.

[26] Sydney Shoemaker, "Causality and Properties," in Peter van Inwagen, ed., *Time and Cause* (Dordrecht: Reidel, 1980), pp. 109–35; Christopher Swoyer, "The Nature of Natural Laws," *Australasian Journal of Philosophy*, LX (1982): 203-23. I shall not here explore the relationship between S/S and property-counterpart theory, but I make the brief suggestion that perhaps some of the insights of S/S can be absorbed into property-counterpart theory if nomic role is taken to be an especially interesting counterpart relation.

I turn now to a brief discussion of two possible challenges. The first is built around the commonly accepted claim that not all coextensive properties are necessarily coextensive. It seems that I must reject this, at least for fundamental properties, once I assert that such properties are world bound.[27] This is parallel to Lewis's rejection of coincident entities. Where we seem to have two objects occupying exactly the same spatiotemporal region but differing in their modal properties, Lewis posits one object with different counterpart relations.[28] Intuitively, the position might look like this. One object might be thought of as, on the one hand, a lump of clay and, on the other, a statue. These two ways of thinking of the object carry with them different weightings for important respects of similarity, yielding different counterparts. Likewise, one property in a given world can be described and thought of in two different ways. The two ways of thinking of the property may well yield two distinct counterpart relations for the one property. The effect would be that under the one description the property will have different modal properties than it does under the other description. The effect of that is that we shall come to think of it as two distinct, though coextensive, properties.

So the counterpart theory for properties allows us to say that there are a property x and a property y such that x and y are contingently coextensive. This is not quite the same as saying that x and y are distinct. But what is our reason for accepting that latter claim? The two most obvious reasons are that we think of the properties in two different ways and that they have different modal properties; and the counterpart theory captures those two truths. It just denies that those truths are sufficient for denying identity. Once we accept counterpart theory, we should be willing to accept contingent identity; and once we accept that, there is no basis for moving from contingent coextensiveness to nonidentity in the actual world. There is no difference here between counterpart theory for individuals and counterpart theory for properties.

---

[27] My thanks to Johanna Seibt.

[28] See, for instance, Lewis's "Counterparts of Persons and Their Bodies," in his *Philosophical Papers* (New York: Oxford, 1983), pp. 47-54. I take a similar line in, for instance, *The Ontology of Physical Objects*.

I believe that this should satisfy our intuitions about most cases that motivate belief in distinct but coextensive properties. But perhaps there are other cases that are not comfortably handled in this counterpart-theoretic way. For instance, if we consider a world in which two properties are coextensive except for a single point at which one is instantiated but the other is not, we may be drawn to the conclusion that, if not for that point, the two properties would have been coextensive but would not have been identical. If such cases motivate us to insist upon the possibility of genuinely distinct coextensive properties, it is easy enough to have them. We can attain this genuine distinctness by simply having distinct property numbers paired with all the same location coordinates within a single world and continue the stipulation that different numbers must always represent different properties.

Initially, this may seem to be more like what we have in mind with respect to all cases that motivate a belief in distinct coextensive properties, but I have my doubts. In the case of genuine distinctness, the coextensive properties are indistinguishable from our point of view. We do not refer to the properties differently, describe them differently, or think of them differently, nor can we stipulate a case in which we do. The effect is that whatever counterpart relations are appropriate to the one property are also appropriate to the other; and if one of the properties has divergent counterpart relations corresponding to two different ways of thinking about it, the other property also does. So these two properties will not differ in their modal properties. Ironically, the case in which there really are distinct coextensive properties is not the sort of case that grounds most of our thinking about coextensive properties.

Let us turn now to a second challenge to my ersatzism. This challenge concerns the possibility of there being no interpreters. In a world with no interpreters, the sets that actually are possible worlds go uninterpreted. In those worlds, then, the sets do not represent. If what is possible in a world *w* depends upon what *worlds* are accessible to *w*, then, since no worlds exist in worlds without interpreters, nothing is possible (or necessary) in those worlds. So if there had been no interpreters, there would have been no modal truths. But that sounds very wrong, and certainly not something that should be a consequence of a theory of possible worlds.[29] The mistake in this challenge is to assume that which sets count as worlds in a world *w* depends on the interpreters in *w*. Even in the actual world, there may, for all I know, be people who have assigned different interpretations to the sets that I have used as my worlds. Certainly, in the actual world, the vast majority of people do not assign any interpretation to those sets. This does not mean that those sets, under the modified Quinean interpretation that I have been discussing, cannot serve to ground the modal discourse of all of these people. What *we* want is some objects that are well suited to the role of possible worlds. Those objects exist even in the worlds with no interpreters; and they are possible worlds because it is *we* who are using them as such. When we say that, if people had never existed, they still might have existed, we are saying something that can be analyzed as: there are worlds in which there are no people and there are other worlds in which there are people and some of the second worlds are accessible to (close enough to) the first worlds.

---

[29] My thanks to Ehring and, independently, Howard Robinson.

What about the utterances of people who exist in worlds in which no one interprets sets as worlds? Well, we should say the same of them as what we say of people in the actual world who have never interpreted sets as possible worlds. What they say can still be analyzed in terms of those possible worlds, even though they may not realize it. In general, speakers need not be able to analyze their utterances in order to be able to utter them appropriately. We do want to be able to say truly that, if no one had ever interpreted sets as possible worlds, then sets would not have represented alternative possibilities. This is true because the sets in a world devoid of interpreters would not, *in that world*, have represented. But this is consistent with saying that in that world it is true that different things could have happened. This latter is a truth because even in that world there are worlds—that is, sets under *our* interpretation—in which other things happen. Assuming an unrestricted accessibility relation, *p* is possible in *w* if and only if *p* is true in some world. It does not matter whether there are any inhabitants of *w* who count the worlds as worlds.

So, overall, I think that linguistic ersatzism is a live option, an option strengthened by its inclusion of a counterpart theory for properties analogous to Lewis's own counterpart theory for individuals. It is by means of this property-counterpart theory that my ersatzism can withstand Lewis's forceful attacks. Property-counterpart theory fits nicely with linguistic ersatzism; if worlds represent properties, instead of containing them, then it is not surprising that which properties are represented is dependent upon our context-sensitive similarity judgments. Admittedly, my ersatzism is still programatic.

BOOK REVIEWS

Towards Justice and Virtue: A Constructive Account of Practical Reasoning. ONORA O'NEILL. New York: Cambridge University Press, 1996. x + 230 p. Cloth $54.95, paper $18.95.

Anyone who takes the history of philosophy seriously must be puzzled by contemporary claims that justice and virtue are opposed. For philosophers both ancient and modern, the relationship between the two is neither antagonistic nor one of mere compatibility; it is symbiotic. Good laws shelter and nurture good characters, without which laws themselves are subject to abuse, misinterpretation, and neglect. Mindful of the disparity between positions past and present, Onora O'Neill explores and critiques the contemporary view; she seeks to develop in its stead a constructivist ethics both sensitive to the concerns that gave birth to contemporary claims and awake to varied and weighty demands for an integrated account of justice and virtue.

O'Neill terms justice-based accounts *universalist*; they appeal to principles that hold across lives, places, and times. Proponents may be perfectionists, consequentialists, deontologists. Contrasting virtue-based or *particularist* accounts span the ranks of communitarians, radical particularists, Wittgensteinians, and Aristotelians. They ground ethical claims not on universal principles, but on community practices and traditions or on the feelings, judgments, or relationships of situated individuals. O'Neill thus attempts to mend a widely accepted rift in contemporary moral philosophy, one that reflects, on her view, a serious error. This is a wide-ranging yet beautifully integrated work that connects many topics of significance in moral theory, weaves together strands of O'Neill's own recent writings, and connects all to concrete moral concerns, from poverty to environmental degradation.

We can trace the divergence of justice and virtue to the intellectual crisis spawned by modernity's deepening suspicion of what had passed for religious and metaphysical certainties. While justice and virtue faced the same foundational challenge, each ultimately answered it in a different way, neither providing any solid grounding for her claims (section 1.4).

If we are to reestablish justice and virtue as mutually supporting, then, we require initial ethical starting points that all persons have reason to accept and a means of moving, with justification all can recognize, to substantive moral conclusions. O'Neill argues that ab-

0022-362X/98/9506/317–21

stract minimal starting points, though appropriate foundations, are alone inadequate to guide our deliberation in concrete cases. We further require a conception of practical reason, itself constructed without reference either to the unvindicated ideals popular among universalists or to the arbitrary norms of actual communities and the personal sentiments and relationships to which particularists of various stripes appeal (sections 2.2-2.3). Addressed to the agent, her conception identifies reasoned action as informed by principles all can follow, where the scope of 'all' varies with context (52).

This conception of practical reason demands that principles of action be *followable*—intelligible to agents and ones they could adopt to guide action (57-58). Noteworthy in O'Neill's elucidation of these two elements are: her defense of abstraction as ubiquitous, familiar to all and friend (not foe) of intelligibility (67-69); her distinction among senses of 'universal' and their respective bearing on charges of rigorism (section 3.3); her challenges to variations on the claim (John McDowell's among them) that abstract universal principles are too empty to guide action (section 3.4). On O'Neill's view, such principles shape action, give it direction, and may be revised based on discoveries made in following the path on which they set us. They do not determine, dictate, or exist immutable for all time (84-85).

To ground and explain the connection between justice and virtue, O'Neill examines the scope, structure, and content of moral principles. Her approach to ethical standing—thus to scope—is practical. Eschewing comprehensive and definitive criteria, she constructs her account by appeal to the assumptions about others that agents' activities presuppose (section 4.2). In particular, if my action assumes that there are others who could affect or be affected by it in ways dependent on their specific and finite powers (if in O'Neill's terms it presupposes plurality, connection and finitude), I must accord them standing as agents, subjects, or both (section 4.3). Founded on assumptions required for effective action and response, this account is designed to block familiar denials of standing from either the insensitivity to context of which justice-based moralities are accused or the bias of traditions or individuals to which virtue-based moralities may fall prey. It prohibits, for example, denial of standing on the ground of another's exaggerated capabilities (for example, her knowledge) or of oneness between agent and another (for example, husband and wife) where these are belied by our actions (section 4.4).

O'Neill's ethical principles display two noteworthy structural features. First, they are not merely hortatory; they set requirements. In identifying prohibited acts, then, they also pick out those which are permitted or demanded. In identifying obligations, they recognize entitlement and vice versa. In short, they offer fuller guidance than hortatory principles could, whether in the realm of justice or of virtue (section 5.1). Further, the principles of the moral theory are *obligation* not *entitlement* based. O'Neill prefers the duty-based moral perspective because it takes account of all moral obligations (those with and without counterpart rights), because it encourages honesty and clarity in allocating and justifying burdens, and because it separates those moral demands we can specify fully only through institutions from those which stand alone (sections 5.2-5.3). An obligation-based morality, in particular, can make room for justice *and* virtue and help us enunciate and develop the relationship between them (sections 5.3-5.4).

Duties of justice are both required of us and owed to specifiable others, who may claim or waive them. For a structure adequate to organize and guide agents' complex lives, some universal principles of justice must be inclusive, their domain roughly cosmopolitan. These will be far too abstract to guide us adequately in more limited contexts (for example, legal, personal, economic). We supplement them with more concrete principles using the thicker concepts of the context at hand (section 6.1). Employing not an abstract notion of liberty or equality, but a *universalizability* requirement, O'Neill argues for universal inclusive principles rejecting any avoidable injury that is either systematic or gratuitous. Such injury might affect agents and their capacities directly or might be indirect, damaging the institutions, relationships, and environments on which they depend (sections 6.3-6.6).

Some virtues are required, but none is owed to any specifiable other; what is owed cannot be claimed or waived and cannot easily be enforced. Again, we need some universal, inclusive principles to organize and guide our lives in broad arenas and varied relationships (184-86). Of the many required virtues, O'Neill considers at length only those she terms *social* (for example, solidarity, sympathy, beneficence). We cannot universalize inclusive principles of systematic indifference and neglect, O'Neill concludes, any more than those of systematic and gratuitous injury. We must express their rejection both in our characters and by developing and nurturing relationships that allow their exercise (section 7.3). In particular, to address vulnerabilities justice ignores or creates, we must reject direct indifference to others, must support connection and social trust, and must actively cherish both natural and human environments (section 7.4).

So many controversial issues are in play here that there is both much to laud and much that critics may question. I sketch one critique. Justice and virtue, which must proceed hand in hand, are not, for O'Neill, ends to be at length achieved, but ones toward which we ever must progress (section 7.6). Her constructivist approach complements this picture; it arms agents with guides both suited to real-world contexts and open to reevaluation and reshaping in light of changed circumstances or new information. Appropriate to such a practical and agent-centered account, O'Neill vigorously defends abstraction; this is, in particular, a vehicle that helps us to set aside factors morally irrelevant to the question at hand, to recognize relevantly like cases, and to treat them similarly (for example, 40, 68-69).

Less fitting is O'Neill's rejection of idealization. This she terms "dangerous" (at least in the realm of moral as opposed to scientific theory) (42). To enhance or idealize starting points, as with John Rawls's assumption of a well-ordered society or Immanuel Kant's of free, equal, and autonomous members of a kingdom of ends, is, she argues, to work from falsehoods.[1] It risks a theory at best inapplicable to human beings in our world and at worst available to justify standards inappropriate to our circumstances. Further, such idealizations often demand justification they never receive (especially 40-44).

But like abstraction, idealization, properly used, serves an important purpose. It directs us to those factors of greatest moral significance, separating these from less basic concerns, giving them priority and allowing us to investigate their connection to matters less fundamental. In complex real-world circumstances, it can help us identify, order, and appropriately respond to interconnected vulnerabilities of the sort that concern O'Neill. Thus, for example, the autonomy of idealized Kantian agents can serve as a guide in addressing complex issues of criminal sentencing or prison reform in a nonideal world. It alerts us to the possibility that a penal system thwarting rather than fostering inmates' ultimate autonomy may run afoul of justice. Where further vulnerabilities in question are less directly connected with individual autonomy, detrimental effects on inmates may warrant additional burdens to taxpayers or reexamination of widespread sympathy for victims who advocate "just deserts."

There is no question that idealization may be misused to great ill effect even by those who simply *neglect* to take into account the differences between real and ideal. Indeed, a theory that idealizes must

---

[1] The Kant example is mine, not O'Neill's.

(as Rawls and Kant concede) provide increasingly less ideal components to connect idealized starting points appropriately to life's complexities. But abstraction also can be misused (as O'Neill herself concedes). I may abstract away morally relevant details of a situation or neglect the fact that details irrelevant to some questions are of enormous moral significance to others (for example, race, gender, religion). Similarly, abstraction's bracketing of various details (for example of agents' economic class) may be as controversial and as much in need of justification as the enhancements or simplifications of idealization. What we require, it would seem, is not abandonment of idealization, but caution in its application and attention to its more controversial assertions.

O'Neill's rejection of idealization, in short, signals a failure to account for some very real difficulties fallible human agents face as they attempt to act, with justice and virtue, in a morally complex world. Like the abstraction she defends, idealization is a tool well suited to a constructivist theory that is attuned to the needs of moral agents and the complexities of our circumstances. The omission is atypical of this otherwise practically sensitive work.

SARAH WILLIAMS HOLTMAN

University of Minnesota/Twin Cities

Outstanding New Philosophy From Cambridge

Responsibility and Control
A Theory of Moral Responsibility
John Martin Fischer and ***Mark Ravizza***

"*This is quite a good book. It presents in a clear way a plausible, general approach to foundational issues about moral responsibility.*"
— Michael Bratman, Stanford University

"Responsibility and Control *is a work of enormous breadth, depth and significance... In many ways, the most systematic discussion of moral responsibility currently available and very likely the best as well.*"
— Jules Coleman, Yale Law School

This book provides a comprehensive, systematic theory of moral responsibility. Fischer and Ravizza explore the conditions under which individuals are morally responsible for actions, omissions, consequences, and emotions.

Cambridge Studies in Philosophy and Law

48055-8 Hardback $59 95

Available in bookstores or from
CAMBRIDGE
UNIVERSITY PRESS
40 West 20th Street, N.Y., NY 10011-4211
Call toll-free 800-872-7423.
Web site: http://www.cup.org
MasterCard/VISA accepted. Prices subject to change.

NEW BOOKS

RESCHER, NICHOLAS: *Objectivity: The Obligations of Impersonal Reason.* Notre Dame: University Press, 1997. ix, 230 p. Cloth $35.00, paper $16.00; and *Predicting the Future: An Introduction to the Theory of Forecasting.* Albany: SUNY, 1997. xiv, 315 p. Paper $21.95; and *Profitable Speculations: Essays on Current Philosophical Themes.* Lanham, MD: Rowman and Littlefield, 1997. ix, 287 p. Cloth $68.00, paper $24.95.

RESNICK, MICHAEL D.: *Mathematics as a Science of Patterns.* New York: Oxford, 1997. xiii, 285 p. Cloth $45.00.

RICHARDSON, HENRY S.: *Practical Reasoning about Final Ends.* New York: Cambridge, 1997. xiv, 326 p. Cloth $54.95, paper $19.95.

ROCHE, MARK WILLIAM: *Tragedy and Comedy: A Systematic Study and Critique of Hegel.* Albany: SUNY, 1997. xvi, 450 p. Paper $21.95.

ROCKMORE, TOM: *Cognition: An Introduction to Hegel's* Phenomenology of Spirit. Berkeley: California UP, 1997. vii, 247 p. Cloth $40.00.

ROSEN, STANLEY: *Plato's* Statesman: *The Web of Politics.* New Haven: Yale, 1997. x, 197 p. Cloth $27.50, paper $16.00.

ROSENBLATT, HELENA: *Rousseau and Geneva: From the* First Discourse *to the* Social Contract, *1749-1762.* New York, Cambridge, 1997. xiv, 298 p. Cloth $59.95.

ROSS, STEVEN DAVID: *The Gift of Truth. Gathering the Good.* Albany: SUNY, 1997. xv, 265 p. Paper $21.95

SCHICK, FREDERIC: *Making Choices: A Recasting of Decision Theory.* New York: Cambridge, 1997. vi, 163 p. Cloth $49.95, Paper $14.95.

SCRIVEN, TAL: *Wrongness, Wisdom, and Wilderness: Toward a Libertarian Theory of Ethics and the Environment.* Albany: SUNY, 1997. ix, 218 p. Paper $20.95.

SELLARS, WILFRID: *Empiricism and the Philosophy of Mind* Cambridge: Harvard, 1997. 181 p. Paper 17.95, Cloth $35.00.

SEN, AMARTYA: *On Economic Inequality.* New York: Oxford, 1997. xii, 260 p. Paper $14.95.

SHAPIRO, STEWART: *Philosophy of Mathematics: Structure and Ontology.* New York: Oxford, 1997. x, 279 p. Cloth $49.95.

SHER, GEORGE: *Beyond Neutrality: Perfectionism and Politics.* New York: Cambridge, 1997. xi, 251 p. Cloth $54.95, paper $18.95.

SHERMAN, NANCY: *Making a Necessity of Virtue.* New York: Cambridge, 1997. xvii, 387 p. Cloth $59.95, Paper $19.95.

SMITH, BARBARA HERRNSTEIN: *Belief and Resistance: Dynamics of Contemporary Intellectual Controversy.* Cambridge: Harvard, 1997. xxv, 221 p. Cloth $37.50 Paper $17.95.

SMOLIN, LEE: *The Life of the Cosmos.* New York: Oxford, 1997. viii, 358 p. Cloth $30.00.

SOLOMON, ROBERT C. and KATHLEEN M. HIGGINS: *A Passion for Wisdom: A Very Brief History of Philosophy.* New York: Oxford, 1997. xviii, 137 p. Paper $19.95.

THE JOURNAL OF PHILOSOPHY
VOLUME XCV, NO. 7, JULY 1998

ON ASCRIBING BELIEFS: CONTENT IN CONTEXT*

When ascribing propositional attitudes, we aim to characterize our subjects in ways that, among other things, enable us to make sense of them in the distinctive way we aim to make sense of *persons;* and there is no making sense of persons, in the relevant sense, without the possibility of assessing their rationality. This means, at the very least, being able to assess their beliefs for logical consistency, to evaluate their inferences for validity, and to identify intersubjective agreements and disagreements between their beliefs and those of other persons—including, importantly, ourselves. If the sentences we use to ascribe beliefs did not systematically enable such assessments, it is not clear in what sense we could legitimately call them *belief* ascriptions.

I THE LOGIC REQUIREMENT

When we reflect on our ordinary practice of belief ascription, we find, I submit, that we aim to capture the logically relevant features of our subjects' beliefs by using, in our reports, content sentences that have, on the occasions of their use, the logical properties that *they* do. In other words, to guarantee that the content sentences of our (oblique or de dicto) belief reports adequately and systematically reflect the logically relevant features of what our subjects believe, we standardly require that the content sentences themselves

---

* An earlier version of this paper was read at a Symposium on Belief Ascription and Opacity at the Eastern Division Meeting of the American Philosophical Association (Atlanta, Georgia; December 1996), where my commentator was David Sosa. The present version has benefited from Sosa's helpful comments as well as Scott Kimbrough's. Related material was presented at the German Philosophical Congress (Leipzig; September 1996), the Maribor Conference on Naturalized Semantics (Slovenia; June 1996), and to the Philosophy Department at the University of Illinois/Urbana-Champaign (March 1996). Discussions with and comments from participants at these other occasions also proved invaluable.

0022-362X/98/9507/323–53

possess, on the occasions of their use, corresponding logical properties. Let us call this the *logic requirement* (LR).[1]

LR constitutes one of two elements that together provide the basis for a natural and compelling semantic (and not merely pragmatic) explanation (and therefore vindication) of the deeply felt intuition of most speakers that the substitution of coreferential expressions in the content sentences of our belief ascriptions is not guaranteed to be truth preserving. The second is the fact that differences in cognitive significance of the sort that typically result from substituting coreferential expressions in sentences—at least such differences as were of concern to Gottlob Frege—are best understood as the manifestation of differences in the *logical* properties of the relevant sentences.[2] In other words, coreferential substitution typically results in a sentence with different logical properties from those of the original—surely a semantic and not a merely pragmatic difference.

We can now explain the substitutional opacity of belief ascriptions—indeed, propositional-attitude ascriptions generally—as a direct consequence of the fact that they are subject to LR, given the more general fact that coreferential substitution may alter the logical properties of a sentence. Insofar as we require, in order for our belief reports to be true, that the logical properties of their content sentences accurately mirror the logical properties of the beliefs being reported, then, since coreferential substitution cannot be guaranteed to preserve the logically relevant features of a sentence, it will come as no surprise that coreferential substitution within these content sentences is not generally truth preserving.

II GLOBAL LOGICAL STRUCTURE

As compelling as the idea is that LR imposes a fundamental semantic constraint on belief ascriptions, providing a satisfactory systematic account of the sorts of logical difference that LR would have us acknowledge turns out not to be as straightforward as one might like. How, after all, are we to understand the differences in logical potential that may result from coreferential substitution? I suggest that we view such differences as manifesting differences in what we may call the *logical structure* of the relevant sentences.[3]

---

[1] Cf. my "Belief, Substitution, and Logical Structure," *Noûs*, xxix (1995): 275-77, where I first labeled this constraint the *logic requirement*. Many philosophers evidence an implicit appreciation of something like LR, though Donald Davidson is the philosopher who has been, perhaps, most explicit in appreciating and emphasizing the importance of the idea that the correctness of attitude ascriptions requires that their content sentences accurately represent the *logical* properties of the attitudes being ascribed.

[2] I argue for this in "Frege, Sense, and Information Content," *Mind*, ci (1992): 767-91

II.1. To see how an appeal to differences in logical structure can enable us to understand better the sorts of difference in logical potential that may result from coreferential substitution—the sorts of difference I claim are responsible for the substitutional opacity of our propositional-attitude ascriptions—consider the following sentences as they might ordinarily be used:

(1) Superman is stronger than Clark Kent.
(2) Superman is stronger than Superman.
(3) There is someone who is stronger than himself.

Though referentially isomorphic, (1) and (2) obviously do not possess the same logical potential. Sentence (1), for example, may not be substituted *salva validitate* for (2) in every valid argument in which the latter appears as premise, nor vice versa: (3) may be validly inferred from (2) but not from (1). Although we may validly infer either (1) or (2) from itself, we may not validly infer (1) from (2), or vice versa. Hence, the conjunction of (1) with the negation of (2), or vice versa, is not logically inconsistent. But we can easily understand facts like these by appealing to a difference in their logical structures.

For expository purposes, it will be convenient to represent logical structure schematically. I shall also use logical schemata more strictly appropriate for a first-order language. For added convenience, I shall allow myself to speak of certain English sentences as instances of various first-order schemata. I expect nothing substantive to hang on this.[4] With this in mind, we can now think of the logical structure of (1) as exhibited by the logical schema:

(1*) $\Sigma\alpha\alpha'$

The logical structure of (2), then, will be best exhibited by the distinct schema:

---

[3] Cf. Hilary Putnam, "Synonymy and the Analysis of Belief Sentences," reprinted in Nathan Salmon and Scott Soames, eds., *Propositions and Attitudes* (New York: Oxford, 1988), pp. 148-58, where, so far as I am aware, something like this suggestion was first made and brought to bear on the issue of substitutivity in belief reports. Putnam's paper was originally published as "Synonymity and the Analysis of Belief Sentences," *Analysis*, XVI (1954): 114-22 All page citations are to the reprinted version.

[4] There is at least one potential drawback to my adopting the schematic approach to representing logical structure. When the schematic approach is understood in too rigid a fashion, it becomes unclear how we can say of two sentences in different languages that they share the same logical structure—that they are, so to speak, instances of the same schemata. I do not take this to be a difficulty of principle. See footnote 7 below for relevant considerations

(2*) $\Sigma\alpha\alpha$

The logical structure of (3) will be best exhibited by the schema:

(3*) $(\exists x)(\Sigma xx)$

The difference in logical structure between (1) and (2), exhibited by the schematic difference between (1*) and (2*), make it clear why they differ in logical potential in the ways just indicated. The argument schemata (A), (B), and (C) are, as ordinarily understood, valid schemata:

(A) $\underline{\Sigma\alpha\alpha}$
$\therefore (\exists x)(\Sigma xx)$
(B) $\underline{\Sigma\alpha\alpha'}$
$\therefore \Sigma\alpha\alpha'$
(C) $\underline{\Sigma\alpha\alpha}$
$\therefore \Sigma\alpha\alpha$

But the argument schemata (D), (E), and (F) clearly are not:

(D) $\underline{\Sigma\alpha\alpha'}$
$\therefore (\exists x)(\Sigma xx)$
(E) $\underline{\Sigma\alpha\alpha'}$
$\therefore \Sigma\alpha\alpha$
(F) $\underline{\Sigma\alpha\alpha}$
$\therefore \Sigma\alpha\alpha'$

Nor indeed is schema (G), as ordinarily understood, a logically inconsistent schema:

(G) $\Sigma\alpha\alpha' \,\&\, \neg\Sigma\alpha\alpha$

II.2. When schematically representing logical structure, it is essential to bear in mind that whether or not some sentence counts as an instance of a given schema depends on the constraints governing substitution within that schema. As logical structure is usually individuated, the constraints on permissible substitutions concern themselves only with the occurrence, or reoccurrence, of expressions *within* a given sentence. I shall call logical structure individuated in this way *local* logical structure. Hence, two sentences count as sharing the same local logical structure just in case there is no schema of which one is an instance and the other is not.

This local (intrasentential) sensitivity to the occurrence and reoccurrence of expressions enables us to distinguish the logical structure of sentence (1) from that of (2) in a way that appears to account nicely for our standard logical assessments concerning them. Unfortunately, appealing to local logical structure fails to enable us to account for precisely parallel differences in the logical potential of referentially isomorphic pairs of sentences that do not contain multiple occurrences of the same expression.[5] For example, our sentence (1) and the referentially isomorphic sentence:

(4) Clark Kent is stronger than Superman.

possess the same *local* logical structure—they may both be viewed as instances of the logical schema represented earlier as:

(1*) $\Sigma\alpha\alpha'$

and there is no schema of which one is an instance and the other is not. Even more problematically, compare the two referentially isomorphic sentences:

(5) Superman flies.
(6) Clark Kent flies.

They also possess the same local logical structure. Both may be viewed as an instance of the schema

(5*) $\Phi\alpha$

Nevertheless, (1) manifestly differs in logical potential from (4), as does (5) from (6), in at least the same ways as (1) differs from (2).

To understand these differences in logical potential by appealing to a difference in logical structure, we shall need a notion of logical structure that is individuated in a way that is sensitive not only to what expressions occur within a particular sentence, but also to what expressions occur in *other* sentences. But, as any student of elementary formal logic is aware, the potential logical relevance of the reoccurrence of expressions in different sentences already requires that we have the notion of an instance of a schema that is sensitive to what gets substituted in other schemata—what W. V. Quine[6] has called a *corresponding* instance. For example, when we say that schema (H) is a valid argument schema:

$$\text{(H)}\quad \begin{array}{l} (\forall x)(\Sigma\alpha x \rightarrow \Sigma\alpha\alpha) \\ \Sigma\alpha\alpha' \\ \hline \therefore \Sigma\alpha\alpha \end{array}$$

meaning thereby that whenever instances of the premise schemata are true, the relevant instance of the conclusion schema is true, we must assume that the instances *correspond*, where this will usually mean that whatever gets substituted at one occurrence of a given schematic letter must also be substituted at all other occurrences throughout the other schemata. Here, in other words, the constraints that govern substitution within schemata, and thereby determine whether or not a given sentence is to count as an instance of some schema, must also concern themselves with what gets substituted in certain other schemata. Once we fix what we are, in a given context, to count as an instance of a particular schema, this automatically constrains what shall count as corresponding instances of other relevant schemata. So, for example, if we were to count (2) as an instance of the schematic conclusion in (H), then we would ordinarily count (1), but *not* (4), as a corresponding instance of (H)'s schematic second premise. Instead, we shall view (4) as a corresponding instance of the distinct logical schema:

---

[5] So far as I know, Soames was the first to point this out, in his "Substitutivity," in Judith Jarvis Thomson, ed., *On Being and Saying: Essays for Richard Cartwright* (Cambridge: MIT, 1987), pp. 99-132.

[6] *Elementary Logic* (New York: Harper and Row, 1965), p. 39.

(4*) $\Sigma\alpha'\alpha$

While (1) and (4) do not differ in their *local* logical structure, they do differ with respect to what I shall call their *global* logical structure.[7]

Appealing to a notion like global logical structure enables us to understand easily the difference in logical potential between (5) and (6). If we view (5) as an instance of the logical schema:

---

[7] We may say, roughly, that two sentences S and S' will have the same global logical structure just in case for all logical schemata Σ and Σ' and any sentence S^*, S and S^* are corresponding instances of Σ and Σ' (respectively) just in case S' and S^* are also corresponding instances of Σ and Σ' (respectively).

Some have worried that this characterization of global logical structure commits me to supposing that sameness of global logical structure amounts to sameness of linguistic form. But, as I have already noted, whether or not a sentence counts as an instance, or corresponding instance, of a schema depends on the principles that constrain substitution within schemata. So the worry would be legitimate only if these principles permit *all* and *only* expressions of the same linguistic form to be substituted for relevant instances of a given schematic letter. While this may be true in the case of standard, fully regimented, first-order languages, there are good reasons to suppose that it is not true for natural languages. Indeed, below I shall argue that it is clearly false for the 'all' case: *sameness* of linguistic form—in particular, sameness of name shape—(even when the expressions corefer) is *insufficient* to guarantee global coinstantiability. In standard first-order languages, the fact that sameness of linguistic form *is* sufficient is, of course, a consequence of the fact that such languages, by design, contain no referentially ambiguous expressions. Although I do not address the issue explicitly here, the 'only' case is also dubious: there may well be contexts in which *difference* of name shape is insufficient to *preclude* global instantiability as well. This might occur, for example, in a context where, for a given speaker, two expressions should be treated as mere stylistic variants—say, the way a name and a nickname might, on occasion, be (or as John McDowell puts it, as differing "merely as objects"—"On the Sense and Reference of a Proper Name," *Mind*, LXXXVI (1977): 159-85, here p. 176).

(5*) $\Phi\alpha$

then we shall typically want to view (6) as a corresponding instance of some appropriately distinct schema, such as:

(6*) $\Phi\alpha'$

So understood, the difference in logical potential between (5) and (6) is transparent. What I am proposing, then, is that whenever the substitution of coreferential proper names within a sentence results in a sentence with a different logical potential—and so, a different semantic content—this is precisely because the substitution fails, in the context in which it is made, to preserve the global logical structure of the original.

II.3. A Fregean, of course, need have no problem with the idea that a difference in the logical potential of two otherwise referentially isomorphic sentences reflects a difference in their logical structure. She will insist, however, that such differences in logical structure are themselves to be *explained* by appealing to differences in the *senses* of counterpart subsentential expressions.[8] In virtue of a difference in their associated senses, the relevant subsentential expressions will make distinct compositional contributions to determining the logical structure—and so the content expressed by—the sentences in which they appear. Explanatory strategies like this, which seek to account for differences in the logical properties of otherwise referentially isomorphic sentences by postulating a difference in the semantic contents of counterpart subsentential expressions, are typically driven by a commitment to an unrestricted principle of compositionality for semantic content.[9] Thus, in cases where the substitution or exchange of coreferential expressions results in a sentence with a different logical potential—and, so, a different logical structure—from the original, it is natural to suppose that there *must be* some distinct, semantically relevant content or value (over and above their reference) associated with the relevant expressions, appeal to which *compositionally explains* the different logical structures.

Sense-theoretic strategies of this sort are committed to viewing the senses of subsentential expressions as possessing an explanatory priority over the logical properties (the logical structure) of the sentences in which those expressions occur. But for an appeal to a difference in the senses of subsentential expressions to afford us a genuine *explanation* of puzzling differences in the logical properties of the sentences in which they occur, the following conditions (at least) must be satisfied. (a) The sense-theorist must provide us with a more substantive account of the notion of *sense*—in particular, an account that specifies what difference in sense amounts to *independently of an appeal to these very logical differences* (or anything that depends upon them).[10] (b) She must show us *how*, on the basis of just such a difference, distinct (coreferential) expressions differ in the compositional contribution they make to the logical properties of the sentences in which they occur. Finally, (c) her sense-based explanation of the logical properties of sentences must support not only our usual assessments of the consistency, inconsistency, and the like of sentences as used by the same person, but also such assessments as they concern sentences used by different people. It must, that is, support our assessments of when, by competently uttering the sentences they do, different people say things that are in agreement (say the same thing), or are merely logically consistent, or are in disagreement and, therefore, logically inconsistent.

---

[8] Whether or not this is a correct interpretation of Frege is not something with which I shall concern myself here.

[9] An unrestricted principle of compositionality will require that of any complex expression E containing an expression n, if the semantic content of some expression n' is the same as (different from) that of n, then the expression E' that results from substituting n' for n in E will possess the same (a different) semantically content as (from) E.

I do not pretend to have a decisive argument showing that nothing worth calling a theory of sense meeting these requirements is possible. For a variety of reasons, however, some of which shall emerge below, I have grown deeply skeptical that any satisfactory account of sense capable of playing this kind of explanatory role is likely to be forthcoming.[11] I have come to believe that the whole strategy of attempting to explain puzzling differences in logical potential by appealing to differences in the compositional contribution of counterpart subsentential expressions is misguided. Consequently, the traditional commitment to unrestricted compositionality—at least insofar as it is understood to commit one to such an explanatory strategy—now strikes me as untenable. I want to urge, therefore, that we try to free ourselves from the feeling that we *need* this kind of compositional explanation of differences in the logical properties of sentences.[12]

---

[10] Cf. Stephen Schiffer's distinct "Intrinsic-Description Constraint" on modes-of-presentation, which may in part be motivated by a similar worry ("The Mode-of-Presentation Problem," in C.A. Anderson and J. Owens, eds., *Propositional Attitudes: The Role of Content in Logic, Language, and Mind* (Stanford: CSLI, 1990), 249-68, see pp. 255ff., and "Belief Ascription," this JOURNAL, LXXXIX, 10 (October 1992): 499-521, see p. 511ff.). But the metaphysical terms in which he articulates his constraint makes it appear vulnerable to criticisms like that of Christopher Peacocke (*A Study of Concepts* (Cambridge: MIT, 1992), pp. 122ff.). The requirements that I propose, however, are requirements on explanatory adequacy, *given a particular explanatory project*. They are, in other words, meant to apply only to the sense theorist who holds that appealing to a difference in the sense of counterpart subsentential expressions actually *explains* differences in the logical properties of referentially isomorphic sentences. To the extent that Peacocke views his theory of concepts as playing *this* kind of explanatory role, I do not believe that his theory satisfies the requirements I have articulated. But it is unclear to me whether, in fact, Peacocke views his theory in this way. McDowell, on the other hand, is admirably clear about the difficulties here and, so, explicitly eschews any such explanatory pretensions for the "modest" (anemic?) notion of sense that he wants to defend (*op. cit.*). But he does still want to defend the theoretical importance of a notion of sense for subsentential expressions. I can find a legitimate place for the notion of sense only at the level of sentences (as used), however.

As I have recommended elsewhere,[13] we do better to think of the content of a sentence, as used on a given occasion, not merely as a function of the contents of its constituent expressions (and their mode of combination), but also, and *additionally*, of its logical structure. What I am suggesting is that the logical properties that a sentence has, on a given occasion of use, are *not* uniquely determined by the fact that it is made up in the way that it is out of words that have, on that occasion, the semantic contents that they do. This is to accord to the notion of logical structure a fundamental status inconsistent with the traditional, sense-theoretic understanding of the explanatory priorities. If I am right about this, then we must reject the idea that semantic content is subject to unrestricted compositionality and accept instead a principle of compositionality that requires—nonvacuously—the preservation of global logical structure. Call this *GLS-restricted compositionality*.

II.4. To accept GLS-restricted compositionality is, in effect, to accept that the logical properties possessed by a sentence, on a given occasion of use, are to a certain extent *autonomous* from the semantic content possessed, on that occasion, by the individual subsentential expressions making it up. They are autonomous in the sense and to the extent that two sentences can, when used, differ in their logical properties—specifically, in their global logical structure—even though they possess the same grammatical structure *and* the counterpart subsentential expressions each possess the same semantic content. I shall refer to this as the *limited autonomy of global logical structure*.

---

[11] A general difficulty for any such approach arises from the fact that while the demand to do justice to our *intra*subjective logical assessments tends, notoriously, to require that senses be individuated in quite fine-grained ways, the demand to do justice to our *inter*subjective logical assessments tends to preclude such fine-grained distinctions.

[12] I share with McDowell "the suspicion that the ultimate source of the desire to see the differences in sense as underlying failures of substitution, and hence as capable of affording genuine explanations of them, is the psychologism about sense, which Frege (officially) renounced" (*op. cit.*, p. 157).

[13] "Belief, Substitution, and Logical Structure."

An important and salutary consequence of accepting the limited autonomy of global logical structure is the liberating effect it has on how finely we are obliged to individuate the semantic content of subsentential expressions. If, for example, we can appeal *directly* to differences in global logical structure, in order to understand differences in logical potential that arise from the substitution of coreferring names, then we are no longer obliged to individuate the semantic content of those names in the incredibly fine-grained way that *would* be necessary in order convincingly to support a *compositional explanation* of the relevant differences in logical potential.[14]

II.5. So far, I have suggested that we ought not to expect—nor should we feel in need of—a sense-compositional explanation of differences in the global logical structure of referentially isomorphic sentences; but I have not offered any alternative to put in its place. If it is not in virtue of a difference in the senses of 'Clark Kent' and 'Superman' that we must view (5) and (6) as correspondingly instancing (5*) and (6*), respectively, then in virtue of what must we do so? The issues here are complicated, involving—among other things—tricky questions about the relation between meaning and use, logic and rational appraisal. Here, I shall try only to gesture in the direction that a more fully adequate answer to this question would take.

The issue, more generally, is what considerations determine whether or not a particular assignment of logical structure to a sentence, or collection of sentences, on a given occasion of their use, is correct? The answer, simply put, is: nothing more nor less than the sorts of consideration that determine in general the correct interpretation of speech. As we already noted, the point of assigning contents to another's utterances—to their propositional attitudes generally—is constitutively bound up with our interest in making sense of them qua rational agents. To identify correctly what another says or believes involves the holistic deployment of a complicated set of context-sensitive, rationality-constrained, interpretive considerations. What beliefs a person has, and gives expression to by means of her assertoric utterances, are fixed in large measure by their place in a necessarily optimally rational pattern of beliefs, desires, and other propositional attitudes. But now, if we carefully reflect on the sorts of consideration to which we appeal when interpreting the speech of another, we find, I submit, that correctly understanding what another says—and, so, correctly identifying the logical properties of the sentences she utters—does not, in any fundamental sense, involve or require first identifying anything worth calling the senses of individual words. Once this fact is noted and adequately appreciated, the fundamental explanatory status I give to logical structure, its limited autonomy from subsentential semantic content, and so my rejection of the explanatory priority typically accorded subsentential content by sense theorists should seem neither surprising nor problematic. My discussion of the examples below will, I hope, further expose the attractions and plausibility of viewing matters in this way.

---

[14] Indeed, this leaves us free, if we like, to identify the semantic content of a proper name with its referent, as is recommended by various *direct-reference theorists*, without, however, having to deny that the differences in cognitive significance which may result from the substitution of coreferential names is of genuine semantic, as opposed to merely pragmatic, significance. I press this advantage in "Belief, Substitution, and Logical Structure." Notice also that by locating the semantically relevant differences at the level of (semi-autonomous) global logical structure, there is no need nor, indeed, any motivation to view uses of belief predicates as involving "unarticulated constituents" in the manner of Mark Crimmins and John Perry, "The Prince and the Phone-Booth: Reporting Puzzling Beliefs," this JOURNAL, LXXXVI, 12 (December 1989): 685-711. Indeed, the kinds of criticism that Schiffer ("Belief Ascription") levels generally against so-called "hidden-indexical" approaches to attitude ascription do not, so far as I can tell, touch our proposal.

III. KRIPKE'S PUZZLE (I)· ASCRIBING BELIEFS TO PIERRE

LR requires that in order correctly to report our own and other people's beliefs, the sentences forming the content clauses of our belief reports must possess, on the occasions of their use, logically relevant features that systematically reflect the logically relevant features of the beliefs we take ourselves to be reporting. We have seen that coreferential substitution fails generally to preserve the logical features of a sentence.[15] We can now see that this is because coreferential substitution generally fails to preserve global logical structure. As a result, belief reports with referentially isomorphic content sentences that differ in their global logical structure will predicate different beliefs of their subjects—beliefs that differ in their potential for logical appraisal. It is precisely this fact, I have claimed, that provides the ultimate semantic basis for the general invalidity of coreferential substitution in the content clauses of attitude ascriptions.

III.1. The substantial attractions of understanding our practice of belief ascription—and specifically the phenomenon of substitutional opacity—along the general lines discussed so far are, I hope, obvious. But additional support for our approach is ready at hand. In "A Puzzle about Belief," Saul Kripke[16] uses his well-known cases involving Pierre and Peter to point out that counterintuitive consequences precisely parallel to those which result when coreferential substitution is permitted can arise in cases that do not in any straightforward way involve violating a restriction on substitution. This suggests—correctly, I believe—that the fundamental source of the counterintuitive consequences is the same in both cases. If this is right, and if our diagnosis of the source of substitutional opacity is correct, then considerations of the sort we used there should enable a parallel diagnosis and resolution of Kripke's two puzzle cases; and this is, indeed, the case.

---

[15] As Benson Mates shows ("Synonymity," in L. Linsky, ed., *Semantics and the Philosophy of Language* (Chicago: University Press, 1950)), and Putnam appreciates (*op cit.*), even co-*intensional* substitution—indeed, even the substitution of *synonyms*—may alter the logical structure of a sentence. Not only does this fact further expose the dim prospects of a helpful compositional-explanatory theory of sense, but the fact that these phenomena *can* easily be accommodated by an approach like that proposed here makes such an approach even more attractive. Cf. my "Belief, Substitution, and Logical Structure."

III.2. We begin with the case of Pierre, a native French speaker, who, because of what he has heard, is disposed sincerely and reflectively to assent to the sentence:

(7) Londres est jolie.

Those of us in a position to translate Pierre's French will ordinarily take this to license our concluding:

(8) Pierre believes that London is pretty.

According to Kripke, our acceptance of (8) is licensed by a combined application of two principles that are presupposed by our ordinary practice of belief ascription. First, given Pierre's assent to (7), an application of a French version of his *disquotation principle* gives us:

(9) Pierre croit que Londres est jolie.

Then, by applying Kripke's *principle of translation* to (9), we get the truth of (8).[17]

---

[16] In Avishai Margalit, ed., *Meaning and Use* (Dordrecht: Reidel, 1979); reprinted in Salmon and Soames, pp. 102-48. All page references are to the latter. From these cases, of course, Kripke wants to conclude that the counterintuitive consequences of strict Millianism cannot be blamed on the fact that it would license such substitutions; for they arise—as in the case of Pierre—even where substitutivity is not an issue. For an earlier, more circumscribed, expression of my own misgivings about the effectiveness of this strategy for defending Millianism, see my "Would a Fregean be Puzzled by Pierre?" *Mind*, xcvii (1988): 99-104.

[17] Kripke's *disquotation principle* (for English) asserts that, if a normal speaker *S* of English sincerely and reflectively assents to a sentence *p*, then '*S* believes that *p*' is true in English, where the sentence replacing '*p*' in both of its occurrences is an appropriate standard sentence of English devoid of indexical or pronominal devices or ambiguities. Corresponding principles apply to other languages. His *principle of translation* asserts that, if a sentence in one language expresses a truth in that language, then a correct translation of it into another language also expresses a truth (in that other language) (*op. cit.*, pp. 112-14).

Eventually, Pierre moves to an ugly part of London, unaware either that there are nicer parts of the city or that the city he now lives in is Londres. Having become bilingual, Pierre is presently disposed sincerely and reflectively to assent to:

(10) London is not pretty.

We take this to license our concluding:

(11) Pierre believes that London is not pretty.

In this case, our license is a direct application of the disquotation principle for English.

Finally, given Pierre's sense of his present surroundings, we may also suppose that he is disposed sincerely and reflectively *not* to assent to the sentence:

(12) London is pretty.

Thus, ordinarily, we would also suppose that we could correctly report:

(13) Pierre does not believe that London is pretty.

According to Kripke, (13) is licensed by a *strengthened*, biconditional form of the disquotation principle (for English).[18] Since we may assume that Pierre has not in any ordinary sense changed his mind about Londres, (8) should remain true.

The problem here should be obvious. In the circumstances, as described, we manifestly have no right to convict Pierre of having inconsistent beliefs. But it would seem that this is precisely what we are forced to do—compare our (8) with our (11). If that were not bad enough, the strengthened disquotation principle evidently forces us to contradict *ourselves*—compare (8) and (13).

But reflection reveals, I submit, that the difficulties that arise in the case of Pierre do so precisely because, as Kripke has presented things, we only *seem* obliged to accept as correct belief attributions whose content sentences in fact fail to represent correctly the logical properties of the beliefs we try to report by using them. As our unwillingness to credit Pierre with logical inconsistency in his beliefs shows, we evidently take his assent to (7) to express a belief the content of which is not logically inconsistent with the content of the belief we take his assent to (10) to express—and so, also, it is a belief that is appropriately different in content from the belief we would take him to manifest were he to have assented to (12).

---

[18] According to this principle, a "normal English speaker who is not reticent will be disposed to sincere reflective assent to '*p*' *if and only if* he believes that *p*" (*op. cit.*, p. 113).

To resolve Kripke's puzzle, then, all we need show is that our ordinary practice of belief attribution does not require, in a case like Pierre's, that we allow as correct such ascriptions as (8), (11), and (13), where—and this is quite important—these are interpreted so as legitimately to license, in the case of (8) and (11), an accusation of logical inconsistency in beliefs to Pierre and, in the case of (8) and (13), an accusation of inconsistency to ourselves. But to suppose that we are so committed is, I suggest, simply to ignore or overlook LR. What we need to show, then, is that one or another of our original ascriptions, (8), (11), or (13)—interpreted in such a way as to underwrite Kripke's puzzle—is *not* in fact licensed. Alternatively, to the extent to which one accepts that the *form of words* exemplified by (8), by (11), and by (13) *are* licensed, we shall want to show that it is a mistake to assign to them interpretations that would underwrite the problematic accusations of inconsistency. Either way, a sensitive appeal to considerations of global logical structure will provide us with precisely what we need.

III.3. Our initial willingness to accept each of (8), (11), and (13) is in large measure, I suggest, an artifact of Kripke's presentation. We are asked to consider the correctness of each attribution independently of any commitment we may have made to the correctness of the others. In effect, we are asked to consider the correctness of each attribution *out of context*. Indeed, the motivation that Kripke offers for accepting each ascription—when taken in isolation from the others—*is* perfectly compelling. But as we shall see—if it is not evident already—sensitivity to global logical structure requires a more holistic and context-sensitive approach to evaluating the correctness of these belief ascriptions.

In order that we may better imagine the kind of circumstances in which we might actually find ourselves making the relevant *set* of ascriptions, let me first fill in Kripke's story a bit. This could, of course, be done in different ways. Suppose that I meet Pierre, knowing that he has moved from France to an unattractive part of London, and I have reason to believe that he is unaware of the more attractive sections of the city. In conversation, he sincerely and reflectively asserts (10), and all the evidence suggests that he would refuse to assent to (12). On this basis, I come to accept (11) and (13)—in virtue, no doubt, of presupposing something like Kripke's strengthened disquotation principle.

As the conversation continues, Pierre learns that I speak some French and so lapses into French himself. But now suppose that at some later point in the conversation, he sincerely and reflectively asserts (7). From this and the earlier English fragment of our conversation, together with whatever other knowledge I might have independently acquired about Pierre, I come to appreciate that Pierre's situation is in all relevant respects just as Kripke describes it.[19] But in a context in which I have already committed myself to (11) and (13), how am I now to report the belief that Pierre expressed with his assertion of (7)? According to Kripke, recall, I would be licensed to accept (8) by applying a French version of the disquotation principle, thus getting:

(9) Pierre croit que Londres est jolie.

and then translating this into English. By the translation principle, if (9) is true, and if (8) is, in this context, a *correct* translation of (9), then (8) must also be true. But, under these circumstances, is (8)—interpreted so as to sustain Kripke's puzzle—a *correct* translation of (9)?

According to LR, if (9) is a *true* belief report—and I shall assume it is—*its* content sentence must correctly reflect the logically relevant features of the belief Pierre expresses by asserting (7). But recall, the belief we take Pierre to express by asserting (7) is a belief *not* logically inconsistent with the belief we take Pierre to express by asserting (10). So, if our translation is to be correct, the content sentence in our translation must not, on the occasion of its use, be logically inconsistent with whatever content sentence we use to report the belief that Pierre expressed by asserting (10). As Kripke would have us interpret it, however, (8) fails to satisfy this requirement. To sustain Kripke's puzzle, the content sentence of (11) (namely, (10)) and the content sentence of (8) (namely, (12)) must respectively instance the logical schemata:

(14) $\neg\Phi\alpha$
(15) $\Phi\alpha$

But corresponding instances of these schemata are logically inconsistent. So, if the content sentence of (8) is taken to instance (15)—as it must be for the puzzle to arise—then (8) fails in this context to be a *correct* translation of (9).

---

[19] It is worth pointing out here—in connection with my remarks in section II.5—that, under these circumstances, my appreciation that Pierre's assertions of (7) and (10) are not inconsistent manifestly does not depend upon or require my identifying distinct senses for 'London' and 'Londres' as he uses them. Indeed, were there any motivation for assigning them distinct senses, it would be *consequent* upon my prior appreciation of the logical facts. What motivation there might be would itself seem to derive almost wholly from the feeling that we still need a sense-compositional explanation of the relevant logical differences. But now, if we were to understand adequately the interpretive considerations grounding our initial appreciation of the logical facts, what would remain to be explained?

Contrary to initial appearances, then, our ordinary practice of attitude ascription, committed as I claim it is to LR, does *not*, under these circumstances, license us to accept (8)—or rather, to be more precise, (8) interpreted so as to sustain Kripke's puzzle. Rather, in this context, we shall require that the content sentence we use to report the belief Pierre expresses by asserting (7) have, on this occasion of its use, a global logical structure that (in an appropriate way) absolves it from logical inconsistency with the content sentence of (11).

That we would in fact recognize and in some way seek to accommodate this demand is manifest, it seems to me, the moment we seriously reflect on how we might actually handle a case like Pierre's. In a discourse context such as we are imagining, where we have already committed ourselves to asserting (11) and (13), we would not ordinarily try to offer (8) as a correct report of the belief Pierre expresses with his assertion of (7)—at least not without further comment. What seems to me far more likely—though this is by no means the only option—is that we would simply appropriate the French name-shape 'Londres'—in effect, thereby, expanding (if only temporarily) our English—and offer instead something like:

(16) Pierre believes that Londres is pretty.

explaining, if need be, that Pierre does not know that Londres and London are one and the same. (In fact, this is precisely what *I* did earlier when, in laying out Kripke's scenario, I said, "Eventually Pierre moves to an ugly part of London and is unaware either that there are nicer parts of the city or that the city he now lives in is *Londres*.")

Given our discussion so far, the motivation we have for doing something like this should be transparent. Our motivation is guided by a complex combination of semantic and pragmatic considerations. For the time being, however, I want to focus on the semantic motivations; the pragmatic issues emerge, it seems to me, more clearly in the case of Peter's beliefs about Paderewski. Our aim, semantically speaking, is to provide a belief report with a content sentence that possesses, on the occasion of its use, logical properties that systematically correspond to the logically relevant features of the belief Pierre expressed with his assertion of (7). On the assumption that we take the content sentence of (11) to instance the logical schema (14), we shall want the content sentence of our present report correspondingly to instance a logical schema that will not render it logically inconsistent with the content sentence of (11). In the imagined discourse context, this is precisely what (16) does. Relative to taking the content sentence of (11) to instance (14), anyone hearing our reports would ordinarily interpret the content sentence of (16) as a corresponding instance of the schema:

(17) $\Phi\alpha'$

thus avoiding any prospect of imputing logical inconsistency to Pierre's beliefs. Moreover, given this way of understanding the logical structure of the content sentence of (16), there is no inconsistency or conflict between (16) and (11). Thus we also avoid committing ourselves to inconsistent attributions.[20]

III.4. If all of this is right, then the puzzling consequences of the Pierre example, as originally presented, are not, as has often been claimed, due to the falsity of one or another of the principles to which Kripke appeals in setting up the example. (I agree with Kripke that, when taken as intended, they appear to be near self-evident truths.) The problem lies not in the content but in the *application* of Kripke's translation principle. Specifically, as we saw, a *correct* translation of (9) requires that the content sentence in our translation—as in the original—accurately represent the logically relevant features of the belief being reported. Given the context with which we are concerned, this requires that we *not* count (8)—interpreted so as to sustain Kripke's puzzle—as a correct translation of (9).[21]

---

[20] Given that we take (14) and (15) to represent the logical structures of the content sentences of attributions (11) and (13), respectively, we are obliged, when reporting the belief Pierre expresses by assenting to (7), to use a content sentence that has a corresponding logical structure like that represented by the schema (17). Although 'Londres is beautiful' satisfies this requirement, I do not mean to rule out other options. Indeed—as we shall see more clearly when we discuss Kripke's Peter/Paderewski example—we *could*, even in the circumstances as described, use as our content sentence the form of words, 'London is pretty'; but *only if*, in using this sentence, we intend—as we ordinarily would not—that it be taken to have a logical structure represented by (17) The obvious potential here for conversational confusion or misunderstanding is doubtless an important pragmatic reason why this possibility would rarely be taken up.

[21] Mark Richard's account of propositional-attitude ascriptions—as developed in his *Propositional Attitudes: An Essay on Thoughts and How We Ascribe Them* (New York: Cambridge, 1990)—also emphasizes the importance that contextual constraints on correct translation (or, as I would prefer to say, correct *interpretation*) have for determining which sentences may appear in the content clauses of true ascriptions. I find myself disagreeing, however, with the details of Richard's implementation of this insight. His idea, roughly, is that for a propositional-attitude ascription to be true, there must be an appropriate (for the context) translation function mapping the RAM of the content sentence onto the RAM of the language-of-thought sentence that figures in the subject's relevant attitude. (RAMs, or *Russellian annotated*

IV. KRIPKE'S PUZZLE (II): ASCRIBING BELIEFS TO PETER

Kripke's puzzle as it concerns Pierre involves two languages; and our resolution of that puzzle involved exposing a restriction on correct translation overlooked in his original presentation. But according to Kripke, the fundamental issue raised by the puzzle does not depend essentially upon the example's involving two languages and, so, evidently, does not depend essentially upon a question of correct translation. To support this claim, Kripke introduces a second example involving Peter's beliefs about Paderewski. The difficulties that arise in this new case are fundamentally analogous to those which arise in the case of Pierre, but here, as Kripke puts it, "[o]nly a single language and single name are involved" (*op. cit.*, p. 266).

IV.1. In this new example, Kripke asks us to imagine that Peter, a native speaker of English, has learned the name 'Paderewski' as the name of a famous Polish pianist while, in a different context, he has also learned the name 'Paderewski' as the name of a Polish national leader and Prime Minister. He does not know, however, that these are one and the same person. We may now suppose that, intending to speak of the person he believes to be a pianist, Peter (sincerely, and so on) asserts (or assents to):

(18) Paderewski has musical talent.

Thus we may conclude:

(19) Peter believes that Paderewski has musical talent.

As Kripke sees it, "If any notion of translation is involved in this example, it is homophonic translation. Only the disquotation principle is used explicitly" (*op. cit.*, p. 266).

---

matrices, are structured entities pairing the expressions making up a sentence with their referents). On the one hand, I find Richard's explicit commitment to cognitive RAMs—and so to the mentalese sentences with which they are correlated—an inappropriately psychologistic commitment with which to saddle a *semantics* of propositional-attitude ascriptions. More to the point, however, in the present context, is the way in which Richard thinks of his translation functions as mapping linguistic names onto mental names. In order to handle puzzle cases, Richard is obliged to individuate mental names (or, more precisely, what he calls the *representational type* of token mental names; see pp. 182ff.) in as fine-grained a way as a sense theorist would be obliged to individuate senses. For—at the level of mentalese—he is, in effect, buying into the sense theorist's strategy of trying to explain differences in sentential content by appeal to differences in corresponding subsentential expressions. But, not that surprisingly perhaps, Richard's conditions for individuating representational types—in particular, what he calls *interior conditions*—fail to meet the adequacy requirements I identified in section II.3). So far as I can tell, Richard attempts to explain differences in representational type by appeal to features of beliefs that presuppose those very differences in representational type (see especially pp. 182-85).

We are also to suppose that Peter is generally skeptical of the musical ability of politicians. Thus, intending to speak of the person he believes to be a Polish statesman, Peter asserts:

(20) Paderewski has no musical talent.

Here, the disquotation principle will evidently license our concluding:

(21) Peter believes that Paderewski has no musical talent.

Thus, we seem committed to attributing logically inconsistent beliefs to Peter. But, under the circumstances, it is no more plausible to conclude that Peter has logically inconsistent beliefs than it was to suppose that Pierre did.[22] Even if Peter were especially logically acute, things would not have turned out differently.[23]

To resolve this puzzle, we need to show that our ordinary practice of belief ascription does not, in a case like Peter's, permit such ascriptions as (8) and (10), where these are interpreted so as to license an accusation of inconsistency to Peter. To the extent to which we accept that contexts exist in which attributions with the *form of words* exemplified by (8) and (10) are permitted—as I do—we need to show that it is possible to assign to them interpretations that capture the relevant logical relations and so fail to underwrite the problematic accusations of inconsistency. In other words, if we are to do justice to what Peter believes, the content sentences of our belief reports must themselves, in the context in which they are offered, not be logically inconsistent. They must, in effect, be corresponding instances of appropriate logical schemata, something like:

(22) $\Phi\alpha$

for the belief Peter expresses by his assertion of (18), and

(23) $\neg\Phi\alpha'$

---

[22] Cf. footnote 19 above Precisely parallel remarks apply to the present case.

[23] Although Kripke does not extend his example in this way, it should be evident that in the conversational context in which Peter asserted (20), he would surely also have been disposed to dissent from an utterance of 'Paderewski has musical talent', if he believed that the person who uttered this sentence intended his use of 'Paderewski' to refer to the person to whom he (Peter) intended to refer when he asserted (20). In this case, the strengthened disquotation principle would license our accepting as true 'It is not the case that Peter believes that Paderewski has musical talent'. As before, then, we apparently end up committed not only to attributing inconsistency to Peter, but to an inconsistency in our own attributions. So the puzzle raised in the Peter/Paderewski case is indeed precisely analogous to that raised in the Pierre/London case.

for the belief he expresses by his assertion of (20). But how are we to do this?

IV.2. When faced with a case like Peter's, a virtually inevitable first reaction is to suggest that though in some uncontroversial sense we and Peter are both speaking English, Peter can and should be taken as speaking a different "dialect" from ours—at least with respect to his use of 'Paderewski'. As one might say: in Peter's idiolect, 'Paderewski' is a homonym *used ambiguously*; though this is not so in ours. One might now suppose that we could simply adopt a modified version of the strategy used in the Pierre case. Perhaps we can extend our English, this time to include a homonym for 'Paderewski', intending thereby to render our 'Paderewski' ambiguous in a way parallel to the ambiguity in Peter's 'Paderewski'. Could we not then intend the occurrence of 'Paderewski' in (19) and the occurrence of 'Paderewski' in (21) to be occurrences of distinct, though homonymous names, thereby securing for our content sentences the correct logical structures?

Although there are some features of this strategy with which I am sympathetic, I find it ultimately to be unsatisfactory. The proposal, as it stands, formulated in terms that have us speak of 'Paderewski' as *ambiguous* in Peter's idiolect, is not only subject to a serious challenge, but, on the face of it, involves assumptions contrary to the *spirit* of our approach.[34] The challenge I have in mind can be usefully approached by considering an important footnote that Kripke appends to the end of his discussion of Peter—footnote number 37. Anticipating a proposal very much like that just sketched, Kripke writes as follows,

> One might argue that Peter and we do speak different dialects, since in Peter's idiolect 'Paderewski' is used ambiguously as a name for a musician and a statesman (even though these are, in fact, the same), while in our language it is used unambiguously for a musician-statesman. The problem then would be whether Peter's dialect can be translated homophonically into our own. Before he hears of 'Paderewski-the-statesman',

---

[34] If the present suggestion *were* truly analogous to the Pierre case, we would have had to assume that in Pierre's idiolect 'London' and 'Londres' also differed in *sense*—and that it was in virtue of this difference that his two beliefs differ in the logically relevant ways that they do. Using 'Londres' in *our* ascriptions would then have to be viewed as our seeking to introduce into our own idiolect a new expression with the same sense that we identified 'Londres' as having in Pierre's. But our proposal in no way depended upon making such an assumption. Indeed, making any such assumption is precisely what we there sought to avoid.

> it would appear that the answer is affirmative for his (then unambiguous) use of 'Paderewski', since he did not differ from anyone who happens to have heard of Paderewski's musical achievements but not of his statesmanship. Similarly for his later use of 'Paderewski', if we ignore his earlier use. The problem is like Pierre's, and is essentially the same whether we describe it in terms of whether Peter satisfies the condition for the disquotational principle to be applicable, or whether homophonic translation of his dialect into our own is legitimate (*op. cit.*, p. 145).

In this passage, Kripke suggests that, if we propose that in Peter's idiolect 'Paderewski' is ambiguous, while not being so in our own, then we face the following dilemma. Are we to say that neither of Peter's two uses of 'Paderewski' can be (homophonically) translated by 'Paderewski' as we use it? Or, rather, should we say that one of his two uses can, but not the other? If the latter is the case, what considerations decide which of his two uses is correctly translatable by our own and which is not? No matter how we try to answer these questions, we seem to get into trouble. But without an answer, the proposal under consideration cannot get off the ground.

On what basis, for example, might we suggest that *neither* Peter's 'Paderewski', as he used it in asserting (18), nor his 'Paderewski', as he used it in asserting (20), may be correctly translated by our 'Paderewski'? Surely, we do not want to propose that they differ from our own merely because the descriptive conditions he associates with each are different from the descriptive conditions we associate with ours. (How, on such a proposal, could we reasonably avoid having to treat every name as possessing a distinct sense in each idiolect? But how then could we do justice to intersubjective agreements and disagreements in beliefs?) In any case, as Kripke effectively points out, despite the fact that *we* know Paderewski as the pianist and statesman, we simply would not ordinarily count 'Paderewski' as used by someone who knew only of Paderewski's musical achievements, though not of his political activities (or vice versa), as differing in meaning from 'Paderewski' as used by us; and so it is far from clear what could justify our treating neither of Peter's uses as translatable by ours. But this also shows why we shall have no satisfactory grounds for choosing *one* of Peter's uses, but not the other, as homophonically translatable by 'Paderewski' as we use it. For any grounds we might offer for choosing one, there will be parallel and equally compelling grounds for choosing the other.

Proposing that 'Paderewski' is ambiguous in Peter's idiolect, while not in our own, suggests that the semantic content of 'Paderewski', as used by us, must be different from the semantic content of 'Paderewski', as used on at least one of the relevant two occasions by Peter. Thus, we find ourselves challenged to search for differences in the semantic content of the names as they occur in Peter's assertions of (18) and (20) by appeal to which we can offer a noncircular explanation of why those sentences, as uttered on different occasions, possess the differential logical properties that they must possess in order to avoid accusing Peter of logical inconsistency. It is just this challenge which we seem unable satisfactorily to meet (cf. II.3). Put into our own terms, Kripke's challenge can be viewed as directed at those who insist on a *compositional explanation* of puzzling differences in logical potential. As I emphasized earlier, however, one of the important and salutary consequences of appreciating the limited autonomy of GLS is precisely that it enables us to avoid the inevitably frustrating need to search for more and more fine-grained contents for individual names in order, thereby, to support such a compositional explanation. As tempting, then, as it may initially appear, the suggestion that the first step in overcoming the puzzling consequences in this case will involve viewing 'Paderewski' as an ambiguous name in Peter's idiolect is, it seems to me, a blind alley.

IV.3. As I have already remarked, we can avoid Kripke's puzzle if the content sentences in the ascriptions we use to report the beliefs that Peter expressed by asserting (18) and (20) correspondingly instance logical schemata like (22) and (23), respectively. Notice, however, that, given the limited autonomy of global logical structure, this will *not* require that the names we choose to use in our two content sentences possess distinct semantic contents. Nor, indeed, does it require that they possess semantic contents distinct from the content possessed by 'Paderewski', as *we* use it! Is there, then, anything to prevent us from accepting that utterances of (19) and (21) can correctly report Peter's beliefs? To be sure, in order to do so, we must interpret them so that their content sentences are taken to instance (22) and (23), respectively. But is there any *semantic* reason why they cannot be so interpreted?

Undoubtedly, a principal source of resistance to appreciating that they can be so interpreted is the fact that precisely the same *name shape* occurs in both ascriptions. For obvious reasons, sentence occurrences isomorphically composed of expressions of the same shape with the same semantic contents will *typically* be interpreted as possessing the same logical potential. Is there, however, any good semantic reason to suppose that they *must* be so interpreted or that in fact they always are?[25]

I submit that just as global logical structure possesses limited autonomy from the semantic content of individual words, so it also possesses a limited autonomy with respect to the orthographic or phonological *shape* of those words. To be sure, the fact that we use the same name shape in both ascriptions may very well lead to conversational confusion or misunderstanding. To avoid this, we may—depending upon the conversational situation—recognize a *pragmatic* need to gloss our different uses, offering, for example, something like, '...Paderewski (thought of by Peter as a statesman but not a musician)...' in one case, and '...Paderewski (thought of by Peter as a great pianist but not a statesman)...' in the other. These parenthetical remarks, however, should *not* be thought of as offering more fine-grained specifications of the *contents* of the particular beliefs we intend to be attributing. Rather, they should be viewed (if proffered at all) as conversational aids aimed at getting our audience to recognize the intended global logical structures of our content sentences. (After all, under the appropriate circumstances, I could have used 'Paderewski' and 'Paderewski′' or 'Paderewski$_1$' and 'Paderewski$_2$' just as effectively.) The parenthetical remarks may, of course, also help to indicate what *other* relevant beliefs Peter may have concerning Paderewski—beliefs, knowledge of which may help one's audience to understand better Peter's cognitive situation.

---

[25] Sentences that contain indexicals provide an immediate counterexample to the general supposition. But even in the present case, where (I am assuming) indexicality is not at play, the supposition is, I believe, mistaken. Nevertheless, it is worth noting that our approach does enable us also to handle, in a simple and straightforward way, belief reports with content sentences in which indexicals do occur. Consider, for example, a situation—like the one John Perry describes (in his "Frege on Demonstratives," *The Philosophical Review*, LXXXVI (1977): 474-97)—in which Bob sees a ship, say, the Queen Elizabeth II, docked at a port, but a large building blocks his view of all but the two ends of the ship. Surely, Bob may fail to believe that the ship he sees to the left of the building is the same ship as the one he sees to the right. In this case, it could turn out that we, knowing that there is only one ship, might be able truly to say of Bob:

(a) Bob believes that *that* [pointing to the ship we see at the right of the building] is an American ship.
(b) Bob believes that *that* [pointing to the ship we see at the left of the building] is not an American ship.
(c) Bob does not believe that *that* [pointing to the ship we see to the left of the building] is an American ship.

Under the circumstances, our intuitions about the case are as follows. The joint truth of belief reports (a) and (b) are plainly insufficient to accuse Bob of having logically inconsistent beliefs. This is so despite the fact that the content sentence in (b) is the syntactic negation of the otherwise extensionally isomorphic content sentence in (a). Nor does the fact that the content sentences of (a) and (c) are referentially isomorphic, together with the fact that they are instances of the same sentence form, in any way preclude the consistency of our report (a) with our negative report (c).

Supporting these intuitions is no problem on our approach. Given the discourse context in which they were offered, the content sentences in each of these reports must surely be taken correspondingly to instance the following logical schemata:

(a*) $\Phi\alpha$
(b*) $\neg\Phi\alpha'$
(c*) $\Phi\alpha'$

Corresponding instances of (a*) and (b*) are not logically inconsistent. So, given LR, belief reports whose content sentences correspondingly instance (a*) and (b*) will not report inconsistent beliefs. Given the logical difference exemplified by the corresponding schemata (a*) and (c*), an ascription whose content sentence instances (a*) will not be inconsistent with a negative ascription whose content sentence correspondingly instances (c*). So, extensional (even intensional) isomorphism together with sameness of sentence form do not generally suffice to guarantee sameness of global logical structure.

IV.4. Given our expanded understanding of the limited autonomy of global logical structure, holding that the occurrences of 'Paderewski' in (18) and (20) possess the same semantic content cannot by itself preclude (18), as uttered by Peter, from possessing a global logical structure represented by (22), while (20), as uttered later by Peter, possesses the global logical structure represented by the corresponding schema (23). This is precisely how we are obliged to interpret these utterances if we hope to get the logical relations between his two assertions correct.

On this way of viewing things, we do not need, when reporting Peter's beliefs, to think of ourselves as "extending" our own English to include any new names. We thereby avoid the difficulty raised by Kripke which we previously noted. What is required is only that, in offering (19) as a report of the belief that Peter expressed through his utterance of (18), we in effect intend our content sentence to instance a schema like (22), while, in offering (21) as a report of the belief Peter expressed through his utterance of (20), we intend our content sentence to be a corresponding instance of schema (23). Given the limited autonomy of global logical structure, there is no reason to suppose that our semantic intentions here cannot be satisfied.[28] Since the content sentences of our ascriptions, on the occasions when they are offered, correspondingly instance (22) and (23), there is no longer any question whether our ascriptions correctly represent the logical features of the beliefs being ascribed. Consequently, Kripke's puzzle does not arise.

---

[28] Once again, our *conversational* intentions may require glosses of the sort we have already mentioned. Of course, even coherently to have the relevant *semantic* intentions will require that we believe that our audience will be able to pick up on those intentions, that is, be able to interpret the content sentences of her ascriptions as possessing the relevant corresponding logical structures. This is one of many places where it may not be possible to draw a theoretically well-motivated line between semantic and pragmatic considerations.

V. RESPECTING INTERSUBJECTIVE LOGICAL RELATIONS IN BELIEF ASCRIPTIONS

So far, our approach offers an efficient and seemingly natural way to understand how our belief ascriptions can do justice to the logical relations that obtain *intra*subjectively between Peter's beliefs. But a fully adequate treatment of propositional-attitude ascription must also do justice to logical relations that hold *inter*subjectively between beliefs—in particular, those logical relations which underwrite intersubjective agreement and disagreement in belief. This very important issue is all too often overlooked in current work on propositional-attitude ascription. So what can *we* say about the logical relations that hold intersubjectively between Peter's beliefs and our own (or some third party who shares our views about Paderewski)? In other words, how, on the present proposal, can we report Peter's beliefs together with our own in such a way as to do justice not only to the logical relations that hold between his beliefs but also to the logical relations that hold between his and our own (or, for that matter, someone else's)?

V.1. Consider the following problem. In reflecting on Peter's epistemic situation, we would ordinarily take ourselves to agree with the belief he expresses through his utterance of:

(18) Paderewski has musical talent.

and to disagree with the belief he expresses through his utterance of:

(20) Paderewski has no musical talent.

In reporting the belief I take to be in agreement with Peter's (18)-belief, I would ordinarily assert something like:

(24) I [also] believe that Paderewski has musical talent.

intending thereby to ascribe to myself a belief with the same content as the belief I correctly ascribe to Peter with my assertion of (19). But it would also be natural for me to report the belief I take to be in disagreement with Peter's (20)-belief in essentially the same way, that is, by asserting something like:

(25) [But] *I* believe that Paderewski has musical talent.

intending thereby to ascribe to myself a belief whose content is logically inconsistent with the content of the belief I correctly ascribe to Peter with my assertion of (21).

But now the following question arises: Is the content of the belief that I report myself as having by my assertion of (24) the same as the content of the belief I report myself as having by my assertion of (25)? Two considerations provide at least prima facie reasons for supposing that it is. First, it seems undeniable that, in this case, we would in some sense take ourselves to base both our agreement and disagreement with Peter on the *same* bit of knowledge, namely, our knowledge that Paderewski—whom we know to be both a pianist and a statesman—was indeed a very talented pianist. In addition, it seems that we have a strong pretheoretic intuition that our use of 'Paderewski' in (24) and (25) is not semantically equivocal.

If, however, we suppose that the content of the belief I report by asserting (24) *is* the same as the content of the belief I report by asserting (25), then the problem becomes obvious; for the content of the belief I report using (24) is supposed to be the *same* as the content of Peter's (18)-belief. The content of that belief, recall, is supposed to be *consistent* with the content of his (20)-belief. So, the content of the belief I use (24) to report must itself be consistent with the content of Peter's (20)-belief. The content of the belief I express by uttering (25), however, is supposed to be *in*consistent with the content of Peter's (20)-belief; for it is precisely in virtue of this that we can be said to disagree. But then the content of the belief I report by using (25) cannot be identical with the content of the belief I reported using (24), for that, as we have just seen, *is* supposed to be consistent with the content of Peter's (20)-belief.

V.2. This, it seems to me, is a very nice problem.[27] To resolve it satisfactorily, we shall have to show how my assertion of (24) may be taken to report a belief whose content differs in a logically relevant way from the content of the belief I report by my assertion of (25)—*and* to do so in a way that plausibly accommodates the intuitions, noted previously. Here is my suggestion.

As we already noted, to hold that I agree with Peter's (18)-belief is to suppose that I have a belief whose content is the same as Peter's.[28] Thus, insofar as (24) is taken correctly to report my agreeing belief, its content sentence must be taken correspondingly to instance:

---

[27] Indeed, this seems to me to be a general problem that any adequate account of belief ascription must be able to resolve.

[28] It is not sufficient for agreement in the sense at issue here that we have beliefs with merely extensionally isomorphic contents. By coming to believe that Superman can fly, one would not, in the relevant sense, thereby count as coming to agree with someone who believes that Clark Kent can fly. There may be a weaker sense of agreement according to which these two may be said to be in agreement, but it is not the sense of agreement at issue here.

(22) $\Phi\alpha$

On the other hand, to hold that I disagree with Peter's (20)-belief is to suppose that I have a belief the content of which is logically inconsistent with Peter's. Thus, insofar as (25) is taken correctly to report my disagreeing belief, its content sentence must be taken correspondingly to instance:

(26) $\Phi\alpha'$

which is, of course, the un-negated schema corresponding to:

(23) $\neg\Phi\alpha'$

Is there any semantic reason to suppose that my utterances of (24) and (25) cannot satisfy these requirements? I suggest not; for given our expanded understanding of the limited autonomy of global logical structure, and assuming there are, in the context, no pragmatic impediments, I may perfectly well assert (24), intending that the content sentence correspondingly instance schema (22), and assert (25), intending that its content sentence correspondingly instance (26)—despite the fact that 'Paderewski' as it occurs in the content sentence of (24) possesses the same semantic content as 'Paderewski' as it occurs in the content sentence of (25).

V.3. But now what about the intuition, mentioned earlier, that we in some sense based both our agreement and disagreement with Peter on the same bit of knowledge? If the content of the belief I report with (24) is distinct from the content of the belief I report with (25), then so is the content of any corresponding knowledge. Fair enough—but doing justice to what is valid in the intuition does not require that we identify the content of the knowledge on which we *base* the relevant agreement and disagreement, with the very contents of the beliefs reported by (24) and (25)—even though it is precisely in virtue of the logical features of those contents that we can be said to agree and disagree at all.

Suppose that, in the context in which I offer (24) and (25), I am asked what is it I take myself to know that leads me, respectively, to agree with the belief Peter expresses by his assertion of (18) and to disagree with the belief that he expresses by his assertion of (20). In response, I would very likely say something like the following: "Well, I know that Paderewski is a statesman, and I also know that he is a very talented pianist. Since one plainly cannot be a talented pianist without having considerable musical talent, it is obvious to me that *Paderewski has musical talent.*" But now, how—you might ask—are we to understand the relation of the belief I here report by saying:

(27) It is obvious to me that Paderewski has musical talent.

to the beliefs I report by asserting (24) and (25)—which themselves, recall, were offered in a context in which I report Peter's beliefs using (19) and (21)?

There are, it seems to me, different responses that I might make; and, depending upon the circumstances, there may be no fact of the matter on the basis of which to choose between them. Under the present circumstances, though, perhaps the most natural way of viewing the matter is to understand the content sentence of (27) as having a global logical structure in virtue of which it, too, would count as a corresponding instance of:

(22) $\Phi\alpha$

This makes the basis of my agreement with Peter vis-à-vis his (18)-belief immediately evident. But now, how do I get from (27) to (25), which has a global logical structure in virtue of which it counts as a corresponding instance of:

(26) $\Phi\alpha'$

thereby directly expressing my disagreement with Peter's (20)-belief? But what is the problem? Under the circumstances, it is clear that I believe that the person about whom Peter is speaking when he asserts (20) is the same person as the one I believe to have musical talent. Or, to put this in a way more directly relevant to our present concern, it is clear that:

(28) I believe that Paderewski [thought of by Peter as a statesman and not a musician] is identical with Paderewski [thought of by Peter as a great pianist but not a statesman].

where the content sentence in (28) is to be understood, in this context, as having a global logical structure in virtue of which it may be taken as a *corresponding* instance of the logical schema:

(28*) $\alpha=\alpha'$

Since it is obvious that I have something like this latter belief, we can now easily see how the belief I report by asserting (27) may indeed serve as the basis not only for my agreement with Peter's (18)-belief, but for my disagreement with Peter's (20)-belief. The belief I report by uttering (25)—the belief in virtue of which I count as disagreeing with Peter—is a straightforward consequence of the beliefs I would report by uttering (27) and (28). This, it seems to me, captures precisely as much as is correct about our sense that both our agreement with Peter and our disagreement with Peter are based on the same knowledge. This being the case, neither that intuition, nor the supposition that 'Paderewski' is not used equivocally in (24) and (25) requires that we take the content of the belief I report myself as having by asserting (24) to be the same as the content of the belief I report myself as having by asserting (25). So our proposal not only offers an efficient and natural way to understand how our belief ascriptions can do justice to logical relations that obtain intrasubjectively, but also enables us to understand how they can do justice to intersubjective logical relations.

VI. CONCLUSION

I have been defending the claim that whether or not a given belief report is true depends upon whether the global logical structure of the content sentence, on the occasion of its use, accurately reflects the logically relevant features of the belief being ascribed. This, as we have seen, depends in subtle ways upon the total circumstance of the report. To commit ourselves, in a given conversational context, to a particular belief report is in part to commit ourselves to an assignment of logical structure to the content sentence in our report. We, so to speak, view the sentence as an instance of a particular logical schema. Consequently, we may accept as correct only those belief reports the content sentences of which we may, in that context, legitimately view as corresponding instances of logical schemata that support our assessments of the logical relations that hold between our subject's beliefs and between his and our own or another's.

I also argued that we should give up the standard sense-theoretic strategy of attempting to explain compositionally the logical properties that a sentence has, on a given occasion of use, in terms of the senses possessed, on that occasion, by the expressions making up the sentence together with their mode of combination. We do better to accord a more fundamental explanatory status to the logical properties of the sentence. In particular, we should recognize that the logical properties that sentences have, on the occasions of their use, possess a limited autonomy from the semantic (indeed, even the orthographic or phonological) properties of their constituent expressions and their grammatical structure.

According to our proposal, then, when we commit ourselves, in a given context, to a particular belief report, we in effect fix, relative to that context, what shall and shall not count as relevant corresponding instances. What, more precisely, does this involve? Although I hope that our examination of the various examples we have discussed help to make this clear enough at a preliminary level, I do not take myself here to have provided anything like a complete answer to this question. A more precise account of when, for example, two expressions may or may not count in a given context as globally coinstantiable will depend upon settling a range of tricky and difficult issues concerning the relation between semantics and pragmatics, a discussion of which, I fear, would have taken us too far afield. Rather, in the foregoing—apart from defending LR and the consequent need, when assessing belief reports, to appeal to the global logical structure of their content sentences—I have been concerned primarily with the *negative* task of pointing out what is *not sufficient* for two content sentences to share the same global logical structure (or to count as corresponding instances of relevant logical schemata). As our discussion has shown, a range of puzzles about belief ascription arise precisely when, in effect, assumptions are made (explicitly or implicitly) about what *is* sufficient that are, in fact, too weak to guarantee distinctions in the logical properties of content sentences of the sort needed to sustain LR.

What our investigation *has* shown us, then, is that, in order for two expressions to be globally coinstantiable (in a given context), it is *not* sufficient that they possess the same semantic content. Moreover, and perhaps more radically, we have seen that in order for two expressions to be globally coninstantiable (in a given context) it may also be *in*sufficient that they possess (in addition to sameness of semantic content) the same orthographic or phonological shape. Global logical structure possesses a limited autonomy with respect to the orthographic or phonological shape of expressions as well.[30]

The deep pretheoretic attraction of accepting LR represents a fundamental semantic requirement on belief reports is, it seems to me, extremely difficult to dismiss. As we now see, violations of LR lay at the heart of most if not all substitutivity puzzles concerning belief reports as well as puzzles like Kripke's that do not, at least in any straightforward way, depend upon issues of substitutivity. If the considerations we have advanced above are on the right track, then satisfying LR requires an appeal to the global logical structure of the content sentences in our belief reports, and this in turn requires (assuming, as I do, the bankruptcy of compositional-explanatory sense theories) that we abandon any commitment to a simple principle of unrestricted compositionality for semantic content. Accepting this provides not only the means for a unified and detailed diagnosis of the sources of our puzzlement in a range of familiar puzzle cases involving belief reports, but also with the means for a natural and effective resolution of those puzzles. These advantages are, it seems to me, no easier to dismiss than LR itself. Although there are indeed questions that remain unanswered and issues unresolved, I submit that the present proposal offers a uniquely promising strategy for approaching and understanding the semantics of our belief reports.

---

[30] As I suggested already in footnote 7, it is also doubtful whether it is sufficient for two co-contentful expressions to *fail* to be globally co-instantiable that they possess *different* orthographic or phonological shapes.

WILLIAM W. TASCHEK

Ohio State University/Columbus

TIME AND MIND*

Mind, it has recently been argued,[1] is a thoroughly temporal phenomenon: so temporal, indeed, as to defy description and analysis using the traditional computational tools of cognitive-scientific understanding. The proper explanatory tools, so the suggestion goes, are instead the geometric constructs and differential equations of dynamical systems theory. I consider various aspects of the putative temporal challenge to computational understanding, and show that the root problem turns on the presence of a certain kind of causal web: a web that involves multiple components (both inner and outer) linked by chains of continuous and reciprocal causal influence. There is, however, no compelling route from such facts about causal and temporal complexity to the radical anticomputationalist conclusion, because, interactive complexities notwithstanding, the computational approach provides a *kind* of explanatory understanding that cannot (I suggest) be recreated using the alternative resources of pure dynamical systems theory. In particular, it provides a means of mapping information flow onto causal structure—a mapping that is crucial to understanding the distinctive kinds of flexibility and control characteristic of truly mindful engagements with the world. Where we confront especially complex interactive causal webs, however, it does indeed become harder to isolate the syntactic vehicles required by the computational approach. Dynamical systems theory, I conclude, may play a vital role in recovering such vehicles from the burgeoning mass of real-time interactive complexity.

---

* I would like to thank Tim van Gelder, Randy Beer, Esther Thelen, and Michael Wheeler for numerous discussions and exchanges on all these topics. I would also like to thank friends, audiences, and colleagues at the University of Sussex, the Santa Fe Institute, and the Philosophy-Neuroscience-Psychology Program at Washington University/St. Louis.

[1] See Tim van Gelder, "What Might Cognition Be, If Not Computation?" this JOURNAL, XCII, 7 (July 1995): 345-81; van Gelder and Robert F. Port, "It's About Time: An Overview of the Dynamical Approach to Cognition," in Port and van Gelder, eds., *Mind as Motion. Explorations in the Dynamics of Cognition* (Cambridge: MIT, 1995), pp 1-44. Related arguments are found in Esther Thelen and Linda B. Smith, *A Dynamic Systems Approach to the Development of Cognition and Action* (Cambridge: MIT, 1994); J.A. Scott Kelso, *Dynamic Patterns* (Cambridge: MIT, 1995); Francisco J Varela, Evan Thompson, and Eleanor Rosch, *The Embodied Mind* (Cambridge: MIT, 1991); and Michael Wheeler, "From Activation to Activity," *Artificial Intelligence and the Simulation of Behavior (AISB) Quarterly*, LXXXVII (1994): 36-42.

0022-362X/98/9508/354–76

I. TIME, CAUSE, AND COMPLEXITY

> The heart of the problem is time. Cognitive processes and their context unfold continuously and simultaneously in real time. Computational models specify a discrete sequence of static internal states in arbitrary "step" time (t1, t2, and so on). Imposing the latter onto the former is like wearing shoes on your hands. You can do it, but gloves fit a whole lot better.[2]

Time, it seems, is the skeleton in the cognitive-scientific closet. A recent spate of criticisms (see footnote 1) of the bedrock explanatory apparatus of cognitive science all focus, in varying degrees, on the need to factor timing into acceptable explanations of cognitive phenomena.

The basic argument[3] is quite simple. It consists of three observations and a conclusion. The observations are:

(1) Cognitive processes all take place in real time.
(2) The timing really matters.
(3) Computational models ignore real timing in favor of the artificial notion of a simple sequence of states.
(4) The conclusion is that cognitive phenomena should be analyzed using the temporally richer framework of dynamical systems theory.

Observation (1) is certainly true, at least as a matter of fact. Observation (2) looks plausible enough. As Tim van Gelder and Robert F. Port point out, "the timing of any particular operation must respect the rate at which other cognitive, bodily and environmental processes are taking place" (*op. cit.*, p. 19). There cannot be much doubt about observation (3) either. Certainly, the formal theory of computation makes no concessions to real timing. It concentrates entirely on the sequence of (discrete) states.[4]

Yet for all that, the basic argument still looks too quick to be convincing. One immediate cause for concern is that it would seem to apply equally well to real-world computer systems. Such systems depend on processes that occur in real time, and the timing of events really matters. Programs that must interact with other programs and

[2] Van Gelder and Port, p. 2.

[3] Van Gelder and Port, pp. 18-23; van Gelder, pp 354, 379.

[4] See, for example, Alan Turing, "On Computable Numbers, with an Application to the Entscheidung Problem," *Proceedings of the London Mathematical Society*, II, 42 (1936): 230-65; Aaron Sloman, "The Mind as a Control System," in Christopher Hookway and Donald Peterson, eds., *Philosophy and Cognitive Science: Royal Institute of Philosophy Supplement 34* (New York: Cambridge, 1993), pp. 69-110; and Marco Giunti, "Dynamical Models of Cognition," in Port and van Gelder, eds., pp. 549-71.

peripherals (printers, scanners, and the like) must likewise respect the rate at which various processes are taking place. Yet it would seem premature to conclude that a dynamical understanding is preferable to a standard computational one in such everyday cases.

The real argument, then, must be a little more complex; and so it is. It depends on two additional factors: first, the continuous nature of the relevant processes; second, the complexity of the causal webs in which such processes participate.

A continuous process is one in which the time series of explanatorily relevant substates cannot be reduced to a sequence of discrete states with jumps in between, but instead requires a genuine continuum of states. This is not the case in conventional computers, which are designed carefully so that their proper functioning is explicable in terms of sequences of discrete states. Here, then, is one potential difference between the case of ordinary computation and biological cognition. Perhaps biological processing makes real use of a continuum of possible states. If so, then any formal computational model must leave out something of theoretical importance, whereas a dynamical model (as we shall see) need not.

More important than this, however, is a nexus of considerations concerning the *type* of causal web in which various states and processes may participate; for it is here that we discern the most potent challenge that temporal considerations pose to traditional forms of explanations and analysis. The challenge is to account for cases of what I elsewhere[5] term *continuous reciprocal causation* (CRC).

CRC occurs when some system *S* is both continuously affecting and simultaneously being affected by activity in some other system *O*. Internally, we may well confront such causal complexity in the brain, since many neural areas are linked by both feedback and feedforward pathways.[6] Worse still, we may find processes of CRC that crisscross brain, body, and local environment. Think of a dancer whose bodily orientation is continuously affecting and being affected by her neural states, and whose movements are also influencing those of her partner, to whom she is continuously responding. Imagine playing improvised jazz in a small combo. Each musician's playing is influencing and being influenced by everyone else. CRC looks, in fact, to pervade the field of natural adaptive intelligence. The delicate dance of predator and prey, or of mating animals, exhibits the same complex causal structure. Historically, CRC looms large in cybernetics[7] and is also wonderfully described by the phenomenologist Maurice Merleau-Ponty[8]:

---

[5] In my *Being There: Putting Brain, Body, and World Together Again* (Cambridge: MIT, 1997), ch. 8.

[6] For example, David Van Essen and Jack Gallant, "Neural Mechanisms of Form and Motion Processing in the Primate Visual System," *Neuron*, XIII (1994): 1-10.

> When my hand follows each effort of a struggling animal while holding an instrument for capturing it, it is clear that each of my movements responds to an external stimulation; but it is also clear that these stimulations could not be received without the movements by which I expose my receptors to their influence...the properties of the object and the intentions of the subject are not only intermingled; they also constitute a new whole (*ibid.*, p. 13).

Here, the motions of my hands are continuously responsive to those of the struggling animal, while the animal's wrigglings are continuously shaped by the motions of my hand.[9]

Van Gelder makes this sort of interactive complexity the centerpiece of an extended argument aimed to reveal the shortcomings of computational models of the mind (*op. cit.*). The key move in the argument is the suggestion that agent-environment interactions may involve such a complex and continuous process of "give and take" that the idea of information entering the agent through perception, giving rise to an internal representation and issuing in an action (and that cycle repeating over and over), is just too simple to do it justice. The kinds of agent-environment coupling characteristic of our cognitive engagements with the world are thus said to "*transcend* representation" (*op. cit.*, p. 381). To illustrate this central theme, van Gelder offers a simple, noncognitive illustration: the operation of the Watt (or centrifugal) Governor.

The task of this device (invented by James Watt in the late eighteenth century) is to keep constant the speed of a flywheel to which some machinery is connected. There is a tendency for the speed to fluctuate due to varying steam pressures and workloads. To smooth things out, the amount of steam entering the pistons is controlled by a throttle valve. How might such control be achieved?

One solution (which van Gelder describes as the computational solution) would involve a sequence of steps and measurements. For example, we might program a device to measure the speed of the flywheel, compare this to some desired speed, measure the steam pressure, calculate any change in pressure needed to maintain the desired speed, adjust the throttle valve accordingly, and then begin the whole sequence anew (*op. cit.*, p. 348). What makes this kind of solution computational, van Gelder suggests, is a complex of familiar features. The most important one is representation: the device measures the speed of the flywheel, creates a token that stands for the speed, and performs numerous operations (comparisons, and so on) on this and other representations. These operations are discrete and occur in a set sequence which then repeats itself. The sequence involves a perception (measurement)-computation-action cycle in which the environment is probed, internal representations created, computations performed, and an action selected. The overall device is, in addition, functionally articulated and homuncular in construction; for it involves a division of the problem into distinct parts, each of which is dealt with independently, with the various parts linked by acts of communication (in which *x* tells *y* the value of *z*, and so on). The features distinctive of the computational governor are thus (1) the use of internal representations and symbols; (2) the use of computational operations that alter and transform those representations; (3) the presence of a well-defined perception-computation-action cycle (what van Gelder calls ***sequential and cyclic operation***); and (4) the susceptibility to information-processing decomposition (what van Gelder calls ***homuncularity***).

---

[7] W. Ross Ashby, *Introduction to Cybernetics* (New York: Wiley, 1956), p. 54.

[8] *The Structure of Behavior*, Alden L. Fisher, trans. (Boston: Beacon, 1963), p. 13; also pp. 148, 168

[9] This complex interactive dance is also stressed by Varela, Thompson, and Rosch and is central to Varela's notion of cognition as "enaction."

Now, for the second solution, the one discovered by James Watt himself: gear a vertical spindle into the flywheel and attach two hinged arms to the spindle; to the end of each arm, attach a metal ball; link the arms to the throttle valve so that the higher the arms swing out, the less steam is allowed through. As the spindle turns, centrifugal force causes the arms to fly out. The faster it turns, the higher the arms fly out. But this now reduces steam flow, causing the engine to slow down and the arms to fall. This, of course, opens the valve and allows more steam to flow. By clever calibration this centrifugal governor can be set-up so as to maintain engine speed smoothly despite wide variations in pressure, workload, and so on.[10]

Having presented the two governors, van Gelder devotes much energy to defending the claim that the centrifugal governor, unlike the computational one, does not rely on representation or computation. Rather than rehearse these considerations,[11] I propose to focus on just one point. The deepest reason for supposing that the Watt governor is not representational, we are told, is that the relationship between the angle of the arm and the engine speed exhibits a special complexity that the framework of representations, computations, and information-processing simply cannot capture. It is a complexity which arises out of the process of CRC itself and which is better captured using the alternative apparatus of dynamical systems theory. The key idea is that although there is a relationship between arm angle and engine speed, it is one that is "much more subtle and complex than the standard notion of representation can handle" (*op. cit.*, p. 353). It is more subtle and complex because the arm angle is continuously modulating the engine speed at the same time as the engine speed is modulating the arm angle. The two quantities are best seen as being co-determined and co-determining—a relationship nicely captured using the dynamical apparatus of coupled differential equations (see section II below). The Watt governor then fails to constitute a *computational* device for two reasons: first, because, on van Gelder's account, computation requires the manipulation of representations (*op. cit.*, p. 353); and, second, because there are no discrete operations in the governing processes and hence no distinct sequence of manipulations to identify with the steps in a computational solution.[12] The Watt governor thus fails to exhibit any of the features associated with the computational solution, and for a single deep reason: the continuous and simultaneous relations of causal influence that obtain among the various factors involved. It is this distinctive kind of causal profile which both invites treatment in terms of an alternative dynamical analysis and which causes problems for the alternative (computational and representational) approach.

---

[10] This story is condensed from van Gelder, pp. 347-50.

[11] For discussion, see Josefa Toribio and my "Doing without Representing?" *Synthese*, CI (1994): 401-31. I propose to grant that the Watt Governor is, in some fairly intuitive sense, a noncomputational device insofar as we do not get our best explanatory grip on its operation by means of its information-processing profile.

Why does CRC cause such problems for standard cognitive-scientific modes of analysis and understanding? The root trouble, it should be clear, involves the search for illuminating problem decompositions; for systems whose causal embedding in the world is characterized by processes of CRC are unusually resistant to the kind of divide-and-conquer[13] approach taken by computational cognitive science. In a divide-and-conquer analysis, a complex phenomenon is explained by displaying a certain kind of articulation in the mechanisms that underlie it. This articulation need not appear at the most fundamental physical level(s), but may instead exist only at the level of a virtual machine implemented in the physical hardware. For example, word-processing software creates a virtual machine whose significant features are quite different from those of the host computer. If we want to understand the distinctive phenomena (operations upon sentences, paragraphs, and so on) supported by such software, we need to display the organization and structure of the higher-level virtual machine; for it is only at this level that we find the groupings and chainings of lower-level, basic-machine states that power the characteristic word-processing profile. The object of divide-and-conquer analysis may thus be either a physical machine or a virtual machine running on top of one. Either way, the analysis proceeds by isolating a number of distinct submechanisms, assigning them specific roles and plotting the chains of causal influence that flow between them. The submechanisms, in the case of computational/information-processing devices, are usually understood as performing operations on semantically interpretable syntactic items, such as numbers, strings, and lists, that encode information about the processed text. The flow of communication between submechanisms then makes sense insofar as it enables a variety of different operations to be performed in some useful sequence: for example, first isolating part of the text (by marking it), then moving it, then rendering it in italics, and so on. The analytic strategy is thus one that turns heavily on (a) the decomposition of the (real or virtual) machine into component states, processes, and submechanisms; and (b) some interpretation of the states and structures such that the flow of causation in the machine can be seen to respect constraints on the flow and manipulation of information.

---

[12] In addition, note that real timing is crucial to these interactions in a way that it *is not* crucial to computational operations. The Watt Governor, unlike the computational one, includes no components whose brief is simply to calculate x and output a result within so many milliseconds (see van Gelder, p. 354). Instead, there is a constant temporal coupling in which change in various components is continuously coordinated in real time. In this sense, real timing matters more in the dynamical solution.

[13] See Daniel C Dennett, *Brainstorms* (Sussex: Harvester, 1978), chapter 7.

CRC poses a clear problem for such a strategy. Although CRC is compatible with basic structural decomposition (distinct parts can be linked by complex reciprocal causal chains), it sits ill with the kind of analysis which assigns specific information-processing roles to components and which interprets those components as effecting a specifiable sequence of operations that solves some problem. In short, it sits ill with the standard idea of an algorithmic depiction of systemic activity; for such descriptions posit a sequence of distinct computational steps or actions that can be seen to solve some problem. The contrast between such a depiction and cases involving CRC is quite marked. As another simple example, consider returning a tennis serve:

> The ball is approaching; you are perceiving its approach, are aware of the other player's movements, are considering the best strategy for the return, and are shifting into position to play the stroke. *All this is happening at the same time.* As you move into place, your perspective on the approaching ball is changing and hence so is activity on your retina and in your visual system...the path of the approaching ball affects which strategy would be best and how you move. *Everything is simultaneously affecting everything else.*[14]

In this example, the relevant structure of events is only very coarsely sequential. To cast the phenomena as a step-wise progression (involving a sequence of discrete states) would clearly do extreme violence to the time coordination of events and processes. The motion of the ball, the tracking (and anticipating) eye movements, and the various gross bodily actions are all unfolding in a kind of phased and interanimated parallelism. To recast such a story in terms of a standard sequential algorithmic description would clearly be to obscure much of importance.

The standard divide-and-conquer computationalist approach, as van Gelder and Port point out, tends to do just that, and as such, is fundamentally ill-equipped for the analysis of such complex phenomena. Parallel computational models (such as connectionist and neural network approaches[15]) are superficially better bets. But even these threaten to fall short in some important respects. For one thing, they lack intrinsic temporality. Nothing in a standard connectionist model fixes the real-time profile of events. Such real-time profiles are, for the connectionist as much as the classicist, relegated to the level of "implementation detail"—a result that is profoundly inimical to the dynamicist's insistence that timing is *cognitively* crucial. For another, both connectionist and classical computational models alike are challenged by cases in which processes of continuous causal influence forge analytic unities that criss-cross brain/body/world boundaries. In the tennis example, for instance, we confront inner cognitive processes that must at some point interface with external (bodily and environmental) events whose temporal structure is crucial to the interaction. (Think of tracking the approaching ball, and factoring in leg motions and bodily reorientations.) If we insist on a computationalist, hence fundamentally atemporal, model of all cognitive processes, then such interfaces present a problem; they must be analyzed in a double-sided fashion, by being shown both to engage successfully the real-time phenomenon and to figure in the more temporally-neutral computational specifications that are supposedly constitutive of the cognitive phenomenon. A common upshot of the pressure to preserve real-time profiles at the interface, however, is a kind of creeping temporal infection. The interface mechanism is depicted in dynamical (real-time involving) terms and this "drives dynamical conceptualizations inward, into the cognitive system itself."[16] This process repeats as the interface is itself required to make contact with more central processes and the real-time dynamics seep further and further into the system. The final result according to van Gelder and Port is a fully temporalized model in which all processes, from input all the way to output, are conceived in dynamical terms. One advantage of the fully dynamical approach is thus that it does not need to deploy different tools to model bodily and environmental processes, on one hand, and more traditionally cognitive ones, on the other. Where inner and outer processes are linked in cycles of CRC, there seems to be no need to draw special boundaries around the inner aspects, nor to use a different vocabulary to describe them.

---

[14] Van Gelder and Port, p. 23.

[15] See James L. McClelland, David E. Rumelhart, and the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Volumes I and II (Cambridge: MIT, 1986); my *Microcognition: Philosophy, Cognitive Science and Parallel Distributed Processing* (Cambridge: MIT, 1989); my *Associative Engines: Connectionism, Concepts and Representational Change* (Cambridge: MIT, 1993); Paul M. Churchland, *The Neurocomputational Perspective* (Cambridge: MIT, 1989).

It is this same deep feature (CRC) which thus inclines the radical dynamicist to question even the traditional divisions between cognitive system, body, and world. If similarly complex causal webs sometimes include neural, bodily, and environmental factors (as in the tennis example), then we seem driven to the use of total-state analyses that include all these factors on an equal footing. This, I believe, must be the reasoning behind bold assertions such as:

> Cognitive processes span the brain, the body and the environment: to understand cognition is to understand the interplay of all three. Inner reasoning processes are *no more essentially cognitive than the skillful execution of coordinated movement or the nature of the environment in which cognition takes place.* The interaction between inner processes and outer world is not peripheral to cognition, it is the very stuff of which cognition is made.[17]

All the truly radical consequences (skepticism concerning the notions of computation and representation and concerning the cognitive/noncognitive boundary itself) thus flow from a single deep source: the idea that intelligent action implicates complex webs of CRC influence. The problem, in short, is *not* (simply) that time matters and that traditional computational approaches leave it out. Instead, the real culprit is a very specific kind of causal complexity: one which impedes the step-wise problem decomposition characteristic (so the argument goes) of a computational approach, and which even fudges the boundary lines between the cognitive system and the environment.

---

[16] Van Gelder and Port, p. 29.

[17] Van Gelder and Port, p. ix (my emphasis). For similar comments, see Thelen and Smith, p. 17; Thelen, "Self-Organization in Developmental Processes: Can Systems Approaches Work?" in Mark Johnson, ed., *Brain Development and Cognition: A Reader* (Cambridge: Blackwell, 1993), pp. 555–91, especially p. 580; Varela, Thompson, and Rosch, pp. 172-80.

To sum up, there are three time-related problems that look to motivate recent anticomputationalist conjectures: first, the simple observation that real timing matters deeply to adaptive success and practical intelligence; second, such timing often involves continuous processes rather than discrete steps; third, inner and outer events are linked in complex causal webs that defy decomposition into temporal (and noncognitive) and cognitive (and nontemporal) components. What we need, it seems, is some framework which builds the temporal facts into the very heart of the analysis, which uses a common vocabulary to describe inner and outer processes, and which is able to capture patterns involving CRC influences among multiple (inner and outer) parts. Such, it transpires, is the profile (and hence the attraction) of dynamical systems theory, as we shall now see.

II. TOTAL-STATE EXPLANATIONS

Dynamical systems theory is a well-established framework for describing and understanding the temporal evolution of complex systems.[18] In a typical explanation, the theorist specifies a set of parameters whose collective evolution is governed by a set of differential equations. Such equations always involve a temporal element, and in this way timing is factored into the heart of the approach. Moreover, such explanations are easily able to span organism and environment. In such cases, the two components are treated as a coupled system in a specific technical sense, namely, the equation describing the evolution of each component contains a term that factors in the other systems current state (technically, the state variables of the first system are the parameters of the second, and vice versa).

Thus, consider two wall-mounted pendulums placed in close proximity on a single wall. The two pendulums will tend (courtesy of vibrations running along the wall) to become swing-synchronized over time. This process admits of an elegant dynamical explanation in which the two pendulums are analyzed as a single coupled system with the motion equation for each one including a term representing the influence of the other's current state.[19]

---

[18] For useful introductions, see Alec Norton, "Dynamics: An Introduction," in Port and van Gelder, eds., pp. 45–68; Kelso; R. Abraham and C. Shaw, *Dynamics: A Visual Introduction, Volumes 1–4* (Santa Cruz, CA: Ariel, 1982).

A useful way to think of this is by imagining two co-evolving state spaces. Each pendulum traces a course through a space of spatial and temporal configurations. But the shape of this space is determined, in part, by the ongoing activity of the other pendulum, which is itself behaving in ways continuously modified by the action of its neighbor. This kind of complex, constant, mutual interaction is, van Gelder and others claim,[20] much closer to the true profile of agent-environment interactions than is the traditional vision of a simple perception-computation-action sequence.

The crucial upshot of the emphasis on constant mutual interaction is a corresponding emphasis on what van Gelder and Port usefully term *total state* (*op. cit.*, p. 14). Because it is assumed that there is widespread and complex interanimation between multiple systemic factors (x influences y and z, and x is itself influenced by y, which also influences z, and so on), the dynamicist chooses to focus on changes in *total system state* over time. The various geometric devices used to put intuitive flesh on the models (trajectories through state spaces populated by attractors, repellors, and so on—see footnote 18) thus reflect motion in a space of possible *overall* system states, with routes and distances defined relative to points, *each of which* assigns a value to *all* the systemic variables and parameters. This emphasis on total state marks one of the deepest contrasts between dynamical and standard computationalist approaches, and it is both a boon and a burden, as we shall see.

A natural question, at about this point, is in what sense a dynamical/total-state story counts as a real explanation of some phenomenon? Are such stories not closer to descriptions than explanations? This is a big topic,[21] but one we need not dwell on here. Suffice it to say that there is no doubt but that a good dynamical systems story is explanatory in at least this sense: it has the power to illuminate counterfactuals—to predict how the system would behave if perturbed in various ways. This is unsurprising, since a dynamical model aims to display the constraints that determine how overall system-state changes over time. What we do not find, however, is anything that looks much like an account of systemic *mechanisms*. Rather, we are shown sets of interlinked parameters, state variables, and so on. The discontent that some cognitive scientists feel when confronted with a dynamical story is thus due to the failure to provide something closer to an engineering blueprint. Dynamical models do not speak directly to the project of how to build a device that would exhibit the target features, and in this they differ from standard computational models.

---

[19] See C. Daniel Salzman and William Newsome, "Neural Mechanisms for Forming a Perceptual Decision," *Science*, CCLXIV (1994): 231-37.

[20] See, for example, van Gelder; Randall Beer and John C. Gallagher, "Evolving Dynamical Neural Networks for Adaptive Behavior," *Adaptive Behavior*, I (1992): 91-122; Wheeler. For discussion, see Fred Keijzer and Sacha Bem, "Behavioral Systems Interpreted as Autonomous Agents and as Coupled Dynamical Systems: A Criticism," *Philosophical Psychology*, IX (1996): 323-46; Clark and Toribio, and Rick Grush and my "Towards a Cognitive Robotics," *Adaptive Behavior* (forthcoming).

[21] For discussion, see my *Being There Putting Brain, Body, and World Together Again*, ch. 6.

For present purposes, however, let us put such reservations aside. Let us use the term *total-state explanation* for any dynamical model that emphasizes the fact that all aspects of a system are changing simultaneously[22] and that therefore invites us to understand the behavior of the system in terms of the possible sequences of changes in total state over time. It is this explanatory strategy which speaks most directly to the needs and worries displayed in section I, since it is this which allows the dynamicist to respect the burgeoning complexity of causal webs in which everything (both inner and outer) is continuously influencing everything else. Such a strategy, for example, allows us to give a satisfying account of the action of the Watt governor using a system of coupled differential equations in which the arm angle (and hence the setting of the throttle valve) appears as a parameter in the description of the engine and in which the engine speed appears as a parameter in the description of the governor. The mathematics of such a system of interlocking differential equations perfectly captures the way in which the two systems engage in their continuous, real-time, and effectively instantaneous dance of mutual codetermining interaction.[23]

The question before us is therefore this: Should the undoubted power and success of dynamical explanations in such cases, combined with the demonstration of similarly complex causal webs criss-crossing agents and environments, lead us to reject the computational model of mind? I next scout some reasons to suppose it should not.

III DYNAMICS AND THE FLOW OF INFORMATION

The deepest problem with the dynamical alternative lies precisely in its treatment of the brain as *just one more factor* in the complex overall web of causal influences. In one sense, this is obviously true: inner and outer factors do conspire to support many kinds of adaptive success. But in another sense, it is either false, or our world view will have to change in some very dramatic fashions indeed; for we do suppose that it is the staggering structural complexity and variability of the brain that is the key to understanding the specifically intelligence-based route to evolutionary success. We do suppose that that route involves the ability, courtesy of complex neural events, to become appraised of *information* concerning our surroundings, and to use that information as a guide to present and future action. If these are not truisms, they are very close to being so. But as soon as we embrace the notion of the brain as the principal seat of information-processing activity, we are already seeing it as fundamentally different from, say, the flow of a river or the activity of a volcano. This is a difference which needs to be reflected in our scientific analysis: a difference which typically *is* reflected when we pursue the kind of information-processing model associated with computational approaches, but which looks to be lost if we treat the brain in exactly the same terms as, say the Watt governor, or the beating of a heart, or the unfolding of a basic chemical reaction.[24]

---

[22] Van Gelder and Port, p. 14.

[23] For further mathematical detail, see van Gelder, pp. 356-57.

The question, in short, is how to do justice to the idea that there is a principled distinction between knowledge-based and merely physical-causal systems. It does not seem likely that the dynamicist will deny that there is a difference (though hints of such a denial are sometimes to be found[25]). But rather than respond by embracing a different vocabulary for the understanding and analysis of brain events (at least as they pertain to cognition), the dynamicist recasts the issue as the explanation of distinctive kinds of behavioral flexibility and hopes to explain that flexibility using the very same apparatus that works for other physical systems, such as the Watt governor.

Such apparatus, however, may not be intrinsically well suited to explaining the particular way neural processes contribute to behavioral flexibility. This is because (1) it is unclear how it can do justice to the fundamental ideas of agency and of information-guided choice; and (2) the emphasis on total state may obscure the kinds of inner structural variation especially characteristic of information-guided control systems.

---

[24] For the last two cases, see Brian Goodwin, *How the Leopard Changed its Spots* (London: Weidenfeld and Nicolson, 1994), p. 60.

[25] For example, van Gelder's comments (p. 358) on tasks that may only initially appear to require "that the system have knowledge of and reason about, its environment," and Thelen and Smith's (p. xix) stress on the brain as a thermodynamic system. By contrast, the dynamicist Kelso (p. 288) sees the key problem as "how *information* is to be conceived in living things, in general, and the brain in particular."

The first point is fairly obvious and has already been alluded to above. There seems to be a (morally and scientifically) crucial distinction between systems that select actions for reasons and on the basis of acquired knowledge, and other (often highly complex) systems that do not display such goal-oriented behaviors. The image of brain, body, and world as a single, densely coupled system threatens to eliminate the idea of purposive agency, unless it is combined with some recognition of the special way goals and knowledge figure in the origination of some of our bodily motions.[26] The computational/information-processing approach provides such recognition by embracing a kind of dual-aspect account in which certain inner states and processes act as the vehicles of specific kinds of knowledge and information. The purely dynamical approach, by contrast, seems committed (at best) to a kind of behavior-based story in which the purposive/nonpurposive distinction is unpacked in terms of such factors as resistance to environmental perturbation.

The second point builds on the first by noting that total-state explanations do not seem to fare well as a means of understanding systems in which complex information flow plays a key role. This is because such systems, as Aaron Sloman has usefully pointed out, typically depend upon multiple, "independently variable, causally interacting substates" (*op.cit.*, p. 80). That is to say, the systems support great behavioral flexibility by being able cheaply to alter the inner flow of information in a wide variety of ways. In a standard computer, for example, we find multiple databases, procedures, and operations. The real power of the device consists in its ability rapidly and cheaply to reconfigure the way these components interact. For systems such as these, the total state model seems curiously unexplanatory. Sloman notes that:

> a typical modern computer can be thought of as having a [total] state represented by a vector giving the bit-values of all the locations in its memory and in its registers, and all processes in the computer can be thought of in terms of the machine's state space. However, in practice, this [total-state explanation] has not proved a useful way for software engineers to think.... Rather, it is generally more useful to think of various persisting sub-components (strings, arrays, trees, networks, databases, stored programs) as having their own changing states which interact with one another (*op. cit.*, p. 81).

---

[26] For a similar argument, see Keijzer and Bem, p. 339.

The dynamicist may suggest that this is an unfair example, since a standard computer will, of course, reward a standard computational analysis. This, however, is to miss the real point, which is that information-based control systems tend to exhibit a kind of complex articulation in which what matters most is the extent to which component processes may be rapidly decoupled and reorganized. This kind of articulation has recently been suggested as a pervasive and powerful feature of real neural processing.[27] The fundamental idea is that large amounts of neural machinery are devoted not to the direct control of action but to the trafficking and routing of information within the brain. The point, for present purposes, is that to the extent that neural control systems exhibit such complex and information-based articulation (into multiple independently variable subsystems), the use of total-state explanations will tend to obscure the important details, such as the various ways in which substate *x* may vary independently of substate *y*, and so on.

The dynamicist may then reply that the dynamical framework really leaves plenty of room for the understanding of such variability. After all, the location in state space can be specified as a vector comprising multiple elements and we may then observe how some elements change while others remain fixed, and so on. This is true. But notice the difference between this kind of dynamical approach and the radical, total-state vision pursued in section II. If, as I suspect, the dynamicist is forced to (a) give an information-based reading of various systemic substates and processes, and (b) to attend as much to the details of the inner flow of information as to the evolution of total state over time, then it is unclear that we still confront a real *alternative* to the computational story. Instead, what we seem to end up with is a (very interesting) hybrid: a kind of dynamical computationalism in which the details of the flow of information are *every bit as important* as the larger scale dynamics, and in which some *local* dynamical features lead a double life as elements in an information-processing economy.

This kind of dynamical computationalism is surely attractive. Indeed, it is the norm in many recent treatments that combine the use of dynamical tools with a straightforward acceptance of the notions of internal representation and of neural computation.[28] Nonetheless, such an accommodation is clearly rejected by those who, like van Gelder, depict the dynamical approach as in some deep sense noncomputational. It is now time to face this characterization head on.

---

[27] For example, the 'gating' hypothesis in Van Essen, Charles Anderson, and Bruno Olshausen, "Dynamic Routing Strategies in Sensory, Motor, and Cognitive Processing," in Christof Koch and Joel Davis, eds., *Large-Scale Neuronal Theories of the Brain* (Cambridge: MIT, 1994), pp. 271-300

[28] See, for example, Paul Smolensky, "On the Proper Treatment of Connectionism," *Behavioral and Brain Sciences*, XI (1988): 1-74; and Jeffrey L. Elman, "Language as a Dynamical System," in Port and van Gelder, eds., pp. 195-226.

IV. COMPUTATION AND CONTENT

Three problems make it hard (but not impossible) to map computational ideas on to the dynamical framework. The problems, isolated in section I, are:

(1) The presence of continuous variations in key physical quantities.
(2) The presence of complex loops of reciprocal causal influence.
(3) The difficulty of mating rich external influence, described in dynamical and noncomputational terms, with a computational account of inner activity.

The mere presence of continuous variation does not, I think, pose any real problem. We may concede that the formal theory of computation (as exemplified in Turing machine computationalism) is defined only for discrete-state machines, and that the feature of digitality is crucial to many classical results in the theory of computability. But there is also a less formal notion of computation (with an equally impressive historical pedigree in early work on so-called analog computation) which is tied to the much more general idea of automated information processing and semantically sensible transitions between representational elements.[29] A common worry about the proposal to thus "allow in" analog computation is that such liberality broadens the notion of computation so greatly as to rob it of explanatory force. The worry is that everything then turns out to be *some* kind of computer and hence the thesis that the brain computes becomes trivial and uninteresting.[30] This result, however, does not follow. For one thing, what has genuine theoretical bite is never the bald claim that such-and-such a physical device computes. Rather, it is the much more substantive claim that such-and-such (a device or subsystem) computes *some specific function*. Thus, even if every object computes *something*, our interest is focused on that special class of computations whose implementations support flexible behavior and reason-guided action.[31] Moreover, given the liberal but intuitive characterization of computation as semantically sensible state transitions between representations, physical systems that do not traffic in representations (for example, the solar system, on any intuitive reading) will not count as computational in any case. The threat of trivialization is thus more apparent than real.[22]

---

[29] For example, Bruce MacLennan, "Field Computation: A Theoretical Framework for Massively Parallel Analog Computation," *University of Tennessee Technical Report CS-90-100*, describes a continuous computational system which is in all other respects similar to a traditional computational system.

[30] See, for example, Sloman, p. 71.

[31] This point is due to David Chalmers, "On Implementing a Computation," *Minds and Machines*, IV (1994): 391-402.

The analog-friendly notion of computation thus shifts much of the burden to the presence or absence of internal representations. This seems sound enough, since even van Gelder asserts that "manipulable representations lie at the heart of the computational picture" (*op. cit.*, p. 351). But there is, alas, a large and complex current debate concerning exactly how we should unpack the notion of internal representation itself.[23] One very plausible constraint, however, is that an internal representation must be a reliably identifiable yet in some sense arbitrary configuration[24] that plays a specifiable semantic role as an intermediary in some problem-solving process; that is, it must be an item, state, or process that we can pick out by physical means but whose functional role is that of *carrying information.*

The presence of continuous variability is clearly compatible with meeting this minimal constraint. But what (moving now to problem number (2)) about the presence of complex loops of reciprocal causal influence? Let us assume (for now) that the problematic loop is fully internal and involves some relation of CRC influence binding the activity of two elements. Would it follow that we could not assign representational roles to the elements? Not at all; it might be, for example, that the coupled activity of the elements is itself the vehicle of some specifiable content. Thus, two oscillators, each of which has a distinct, individual frequency of oscillation, can be joined (by continuous reciprocal coupling) so as to constitute a new higher-level element with a new and uniform frequency of oscillation. Technically, the phase of each oscillator continuously influences the periodic behavior of the other resulting in an oscillation frequency for the coupled system.[25]

---

[22] For further discussion, see Brian Cantwell Smith, *On the Origin of Objects* (Cambridge: MIT, 1996); my *Being There: Putting Brain, Body, and World Together Again*, ch. 8; Sloman; Chalmers; and papers in Stephen Harnad, ed., "What is Computation?" *Minds and Machines*, IV, 4 (1994): 377-488.

[23] See, for example, my *Being There: Putting Brain, Body, and World Together*, ch 8; van Gelder; R. Beer, "A Dynamical Systems Perspective on Environment Agent Interactions," *Artificial Intelligence*, LXXII (1995): 173-215, Wheeler.

[24] This term is to be understood broadly so as to include physical states, virtual states, and temporally extended processes; that is, any identifiable feature or pattern capable of entering into causal relations.

[25] See Port, Fred Cummins, and J. Devin McCauley, "Naive Time, Temporal Patterns and Human Audition," in Port and van Gelder, eds., pp. 339-72, here p. 361.

Such regular coupled behavior could itself be used as a systemic stand-in for some state of affairs. (For example, a somewhat augmented circuit, known as an adaptive oscillator, can entrain its activity so as to replicate the frequency of an input pulse and to continue the rhythm in the absence of the input.[26]) The point, in any case, is just that there is no reason why entire inner loops, involving multiple components in CRC exchange, should not, at times, themselves constitute the internal vehicles of specifiable contents. Such loops would stand revealed as high-level syntactic items whose functional role is to carry, communicate, and transform specific bodies of information; in short, as internal representations implemented in complex dynamics.

Here, then, is a new and fascinating way in which dynamical and computational analyses may proceed hand-in-hand. The dynamical analysis may help identify the complex and temporally extended physical exchanges that act as the *vehicles* of representational content. Traditional computationalism may have been too narrow-minded in its vision of the likely syntactic form of the inner bearers of information and content. Our fascination with the static characters and strings of natural language may have led us to expect simple, local, spatially extended states to function as inner content bearers. Connectionist approaches helped us see beyond that vision, by identifying the content bearers as distributed patterns of activity. But it may take the full firepower of dynamical systems theory to reveal the rich space of possible content bearers: a space that includes very complex temporally and spatially extended processes and interactions.

What, though, of those loops of continuous reciprocal influence that look to criss-cross brain/body/world boundaries? These loops are especially problematic, as we saw in section I, because they threaten to create tightly coupled systems incorporating elements (neural, bodily, and environmental) that traditionally require very different kinds of description and explanation. The problem is especially acute when the external influences are described in richly temporal terms while the inner economy is depicted as purely computational.[27] This is the "mating" problem identified as number (3) on our original list.

Once again, the situation looks a lot worse then it really is. First of all, it is simply not the case that temporality must somehow fade away at the point where a computational story takes over. If, as we just conjectured, temporal features of inner processes can play representational roles (and why not?—temporal duration is a perfectly good nonsemantic feature that nature may employ for information-encoding purposes), then the dynamical characterization may be married[illegible] to the informational and computational accounts. In this way, the temporally-characterized inner states and processes are also given an information-processing gloss, and this opens the door to viewing them as participants in processes of (perhaps analog) computation.

---

[26] See Port, Cummins, and McCauley, pp. 361-63.

[27] See van Gelder and Port.

If it is asked why we should offer this further gloss on the inner events and not on the outer ones, the answer is clear enough. It is that *typically*, the inner events are playing a very different functional role from the outer ones. The difference is two-fold. First, the inner events are acting to control the behavior of a well-defined individual organism, whereas the outer events (again only typically) constitute the sources of variance with which the controller must cope. Second, the control in question is special insofar as it is information-and-goal based. Regarding the first point, we can see that even the Watt governor plays a different functional role from the engine; for it is the *job* of the governor to keep the engine speed constant, whereas there is no plausible teleological story in which the reverse is true. Governors evolved to control engines, and this fact alone marks an important difference that is unaffected by the dense continuous and reciprocal nature of the causal exchanges that bind the governor to the engine.[illegible] The Watt governor, however, is the kind of control system that does not in addition require us to invoke an information-based account of its control strategies. Instead, we gain a full and sufficient understanding of its operation by focusing on physical laws relating physical quantities. This is not the case in control systems that, for example, create and manipulate models of the domain with which they deal, or ones that maintain hierarchies of goals and subgoals, reason about what action to perform next, and so on.

[illegible] Just such a union is pursued in James Crutchfield and Melanie Mitchell, "The Evolution of Emergent Computation," *Proceedings of the National Academy of Science*, XCII (1995): 10742-46, and in Mitchell, Crutchfield, and Peter Hraber, "Evolving Cellular Automata to Perform Computations," *Physica D*, LXXV (1994): 361-91. It also characterizes several papers in Port and van Gelder, eds. Van Gelder's own notion of revisionary representationalism and his discussion of decision field theory (pp. 359-63) show that he is open to the idea of such a union of dynamics and representationalism, though troubled by the time-based considerations highlighted in the present treatment.

[illegible] A parallel argument aimed to distinguish genetic control from environmental backdrop is developed in Kim Sterelny, "Understanding Life: Recent Work in Philosophy of Biology," *British Journal for the Philosophy of Science*, XLVI, 2 (1995): 155-83.

To assert simply that human intelligence involves the use of inner models and representations would, of course, be to beg the question at issue. The point is rather that, even if the leading skeptical consideration (concerning relations of CRC influence that criss-cross brain, body, world boundaries) is correct, this is in itself no reason to suppose that we cannot motivate talk of inner models, representations, computations, and all the rest. In so doing, we opt for a distinctive vocabulary and an explanatory style which highlights the special functional role of certain events as control structures and which reveals the control function as based on the extraction, preservation, and use of specific types of information.[40]

Next recall the argument[41] that the complexities of certain organism/environment interactions may (courtesy of our old friend, CRC) be "so subtle and complex as to defy description in representational terms."[42] One immediate reaction is that even here the case remains unproven. Even in the example of the ongoing tennis game, it is by no means obvious that there are not distinct inner states playing specific representational roles that may be identified by future neuroscientific investigations. Moreover, even if the temporal complexities of the interactions cannot be directly captured using talk of representations and information flow, it does not follow that we should reject such descriptions outright. It is quite possible, for example, that the pure dynamical description is *equally* inadequate. If it highlights physical exchanges at the expense of revealing the fine structure of goals, strategies, and desires that help modulate the play, it, too, will leave out much that is of interest. The point about the possible supra-representational status of the agent/environment relation gains its force from an implicit assumption to the effect that one kind of story (either dynamical or representational) ought to tell us *all* that we want to know about the agent-environment relation in question. I suggest, however, that this assumption is unwarranted. It is perhaps fueled by the use of the Watt-governor illustration in which the dynamical account really does seem to tell us all we need to know. But this is not a demonstration we can trust since the governor does not produce the kinds of highly flexible, rapidly reconfigurable, multiple goal-and-task oriented behaviors so characteristic of cognitive agents. The governing task is thus not *hungry* for the kind of information- and representation-based control strategies that seems especially illuminating in cases where the behavior is much more highly flexible and depends on the maintenance of many bodies of information and complex goal structures. In such cases, it is surely implausible to expect any model that eschews talk of information and representation to satisfy us. Rather, we should pursue a hybrid approach that both tracks the complex interactive dynamics and makes some attempt to reveal the information processing roles of component states and processes.

---

[40] For further argument to this effect, see Sloman; Clark and Grush. Notice, in passing, an important wrinkle. By giving prominence to the notion of a distinctive kind of functional role (namely, control by information-and-representation) we allow that some *external* states and processes might indeed count as (perhaps temporary and transient) parts of the cognitive control system itself. What will determine whether specific external processes and events count as part of an extended cognitive system is not, however, the presence or absence of continuous reciprocal causal interactions. For this, we now see, is orthogonal to the real issue, which is whether the external events participate in the encoding, maintenance, or transformation of information used to control intelligent action. Many external structures play—for humankind—just such a role. Think of abacuses, computers, books, ships compasses, maps, and so on. Such items are selected to function as elements in environmentally extended processes of computation and information-based control. In these cases, the inner and the outer do, I think, form extended computational systems that reward treatment as integrated, information-processing wholes. For further argument, see Edwin Hutchins, *Cognition in the Wild* (Cambridge: MIT, 1995), my *Being There: Putting Brain, Body, and World Together*, chs. 9, 10. It is, of course, a further and extremely vexing question when (if ever) we should count such external structures as *part of the agent* herself—for some discussion, see Chalmers and my "The Extended Mind," *Analysis*, LVIII (1998).7–19.

[41] See van Gelder; and section II above.

[42] Van Gelder, p. 381.

But there is, in any case, one last reason why we should not be overly troubled by the spectre of complex causal chains that crisscross the agent/environment boundary. It is that such complexity is not characteristic of most of the (intuitively) core cases of genuinely *cognitive* phenomena. Both historically and intuitively, the very idea of cognition has been tied to (a) the idea of reasoned behaviors selected in the absence of any local guiding stimulus (as in advance planning, counterfactual reasoning, and so on), and (b) to behaviors that seem to track non-nomic stimulus properties (such as "being a movie ticket"—a property which does not figure in any natural laws but one to which we can reliably respond nevertheless).[43] That is, the idea of cognition has been tied to the idea of behaviors carried out in the absence of any constant, lawful, and reliable signal from the local environment.[44] This is why a light-following robot does not strike us as a paradigmatically cognitive engine, whereas a robot that engages in off-line planning and abstract problem-solving does.

---

[43] See, for example, Jerry A. Fodor, "Why Paramecia Don't Have Mental Representations," in P. French, T. Uehling, and H. Wettstein, eds., *Midwest Studies in Philosophy, Volume X* (Minneapolis: Minnesota UP, 1986), pp. 3–23.

[44] See, for example, the definition of internal representation in John Haugeland, "Mind Embodied and Embedded," in Yu-Houng Houng and Jih-Ching Ho, eds., *Mind and Cognition* (Taipei, Taiwan: Academia Sinica, 1995), pp. 3-38, and discussion in Smith.

It would be unfair to make too much of this, since it is an admirable part of the agenda of the dynamicist to combat the "passive, reflective" model of cognition that has so deeply permeated traditional philosophical and scientific thought. Nonetheless, it is surely fair to comment that it was the project of explaining both environmentally detached reasoning and consistent response to non-nomic stimulus properties that originally led theorists to posit inner models and manipulable representations. When, for example, an agent reasons about an event in advance, the theorist is led to ask what structural surrogates the system is using so as to select appropriate actions in the absence of actual local triggers and cues. When an agent reasons about movie tickets, the theorist is led to ask how a physical system can possibly key its responses to a feature of the world that does not fall (qua that feature) under the laws of physics.

In sum, the worries about temporality and complex causal exchanges look, on closer examination, unable to bear the kind of weight required by recent skeptical arguments. It is by no means clear that pure dynamical stories can do justice, even in principle, to the special requirements imposed by the project of explaining reason-guided action. It is also not clear that the kinds of case to which the skeptical considerations most plausibly apply are really on a par with the cases for which the target notions (of internal representation and computation) were originally invoked.

V. CONCLUSIONS. ACCOMMODATION NOT ELIMINATION

Suppose we accept that traditional computational models *have* tended to screen off a little too much of the real physical and temporal structure of events, and have tended to erect somewhat too firm a barrier between the cognitive system and the rest of the world. One response to these excesses is to reject outright the notion that cognition is a matter of computation, and with it the whole apparatus of internal representation and information-processing theory. But this, I have argued, is a distortive and even dangerous overreaction. It is distortive insofar as there is nothing in the detailed considerations and arguments put forward (concerning the importance of timing, and the possibility of CRC influence) that is in any way inimical to a substantive computational and informational analysis. It is dangerous insofar as it threatens to deprive us of the special kinds of tool and understanding needed to make sense of the special class of information-based control systems.

What is indicated is thus not the rejection of the computational/representational vision so much as the addition of an irreducibly dynamical dimension to the analysis. Such a dimension manifests itself in several ways, including: the use of dynamical tools to recover potential information-bearing vehicles from highly complex webs of causal exchange; the recognition that intrinsically temporal features may play representational and computational roles; and the extension of computational ideas to systems that change continuously in time and that exploit continuous state.

Accommodating this dynamical dimension clearly calls for a substantial broadening of the standard computationalist outlook. But as long as we maintain the distinctive and explanatorily potent physical/informational dual aspect that is at the heart of the computational framework, we confront (I suggest) only refinement and progress, not outright elimination. The future of cognitive science may thus lie in a delicately negotiated union between the familiar framework of computational, representational, and information-processing description, and the challenging and temporally charged project of dynamical analysis. For such a union to succeed, the dynamicist must look beyond total-state explanations, attending as much to the details of mechanism and information flow as to the shape of overall systemic unfolding, while the computationalist must recognize the large and baroque space of physical and temporal processes that nature may plunder for representational gain.

ANDY CLARK

Washington University/St. Louis

THE JOURNAL OF PHILOSOPHY
VOLUME XCV, NO. 8, AUGUST 1998

NAMES AND EXPRESSIONS*

Natural languages allow that distinct expressions may refer to the same individual, but it would be wrong to think that it is never important which one of these is chosen. Sometimes it is of the essence. A sentence containing one expression may differ in truth value from a sentence containing another expression coreferent to it, even if the sentences differ in no other way. Analysis of puzzling facts such as this led Gottlob Frege to propose that associated with an expression, or sign, is not only its reference but also its mode of presentation; distinct expressions, he held, may have the same referent but different modes of presentation. He then argued that failures of substitution are only apparent, arising when expressions refer not to their usual referents but to their modes of presentation. More recent explanations for substitution failure continue to appeal to modes of presentation; on one account, an agent may believe a proposition under one mode of presentation, but disbelieve the same proposition under another mode of presentation. Of course, the term 'mode of presentation' is here used in a way different from Frege's, but there is a functional similarity. One may think of a *mode of presentation* as that in a theory which serves to explain why coreferential expressions are not everywhere intersubstitutible (following, in essence, the usage of Stephen Schiffer[1]). Using the term this way, the identity of the modes of presentation may vary from theory to theory.

---

* We would like to thank Kent Bach, Sylvain Bromberger, Graeme Forbes, Ernest Lepore, Brian Loar, Richard Mendelsohn, Stephen Schiffer, Gila Sher, Christopher Tancredi, and especially Barbara Bevington for valuable comments.

[1] "The Mode-of-Presentation Problem," in C.A. Anderson and J. Owens, eds., *Propositional Attitudes: The Role of Content in Logic, Language and Mind* (Stanford: Center for the Study of Language and Information, 1990), pp. 249-68.

0022-362X/98/9508/377-409

It is apparent that, if an account of substitution failure based on distinct but coreferring expressions having different modes of presentation is to be successful, it will be necessary at some point to provide a criterion of mode identity: Given distinct expressions with the same referent, what features of them are distinct such that they could play the mode-of-presentation role? Frege's answer was that the expressions have distinct senses, and that senses play the mode-of-presentation role. Given the nature of his program, this was not unreasonable; but other answers can be given. For example, if we wish to find distinct features of distinct but coreferential expressions, it is clear that among the candidates to consider will be whatever those features are which distinguish the two expressions in the first place. If, in particular, there are formal syntactic criteria in virtue of which expressions are distinguished, those formal distinctions themselves might play the mode-of-presentation role. Although it is often thought that there are insuperable problems with such an approach,[1] we nonetheless feel that such an answer deserves a hearing. There is, however, an important prerequisite to this undertaking. If an account of *how* syntactic expressions serve the mode-of-presentation role is to be forthcoming, we must first have a formal theory of syntactic expression identity for natural language. For reasons that will become clear directly, such a theory must, among other things, formally distinguish syntactic expressions from names. This latter task is one that must be faced in any event, as a chapter in the theory of syntax; it will be our point of departure. We shall then be in a position to consider the fitness of linguistic expressions, so conceived, for the mode-of-presentation role.

I DISTINGUISHING NAMES AND EXPRESSIONS

In a representational sense, the distinction between syntactic expressions and names has been explicit in the syntactic literature for many years. *Names* are lexical items, members of a "lexicon," a list which also contains the other words of the language. They appear as terminal symbols in syntactic structures (phrase markers), which also contain nonterminal symbols, or *nodes*. These nonterminal symbols, inclusive of their terminal, lexical, contents, are the *syntactic expressions*. When names, the sort of words on which we shall focus throughout, appear in syntactic structures, they are dominated by the *node NP*. Thus, given that 'Max' is a name, there are syntactic expressions of the form '[$_{NP}$ Max]'. The distinction between words and expressions arises in natural language in part because of the fact that, in natural language, the form of a word, its phonological "spelling," does not of itself reveal the syntactic category of the word, and the ordering of a string of words does not of itself reveal the syntactic analysis of the string. In natural languages, category membership reveals these things in a way that phonological spelling alone cannot.

---

[1] We have in mind puzzles such as the "Paderewski" puzzle, initially discussed by Saul Kripke in "A Puzzle about Belief," in Avishai Margalit, ed., *Meaning and Use* (Dordrecht: Reidel, 1979), pp 239-83. We discuss this case, and the related "Londres" puzzle in section IV.

There are certain important differences between the terminal and nonterminal symbols to be noted. The nonterminal symbols—nodes such as NP, VP, and the like—are members of a very small universal inventory. The theory of natural-language syntax, at least ideally, is sufficient to determine their identity, and the linear and hierarchical relationships that may hold among them. With respect to these matters, the theory of syntax is very narrowly defined. The terminal vocabulary is very different in these respects. If we survey the lexicons of individual speakers, even those who would be said to share a language, we find significant variation in the inventory of entries. It would be very rare for two individuals to share a lexicon fully; an individual has some words in his lexicon and not others. This being so, if we are to determine the identity conditions for words and, in particular, names, we should determine the identity conditions on words *as they appear in the lexicon of an individual.* We shall find that the *beliefs* of individuals will be criterial in the statement of those identity conditions on words: two occurrences of words are occurrences of the same word for an individual if and only if that individual believes they are the same.

What criteria, then, does a speaker use to determine whether she is confronting two occurrences of the same word, or occurrences of distinct words? It seems clear that pronunciation can be criterial: a speaker may believe that there are two occurrences of distinct words, and not two occurrences of the same word, if the speaker believes that there is distinctness of pronunciation between the two occurrences. But what sort of distinctness of pronunciation counts? Consider the sorts of distinctions one finds between the voice quality of two individuals: suppose one is young and the other old, or one is female and the other male. Although their pronunciations will be acoustically distinct, and easily discriminable, people do not consider a word said in a male voice to be different from the same word said in a female voice. In general, the sorts of acoustic considerations not considered linguistically relevant in the broad phonetic transcriptions of phonologists are not criterial for the lay person in the individuation of words.

Nevertheless, there are circumstances under which the lay person allows that one word may be pronounced quite differently, where the differences in pronunciation are linguistically relevant. A speaker who grows up speaking one dialect of English, and then moves to an area in which an unfamiliar dialect of English is spoken, will learn a set of correspondences (for example, when they say 'Noo Yawk' they mean 'New York') and then might come to be fluent in both dialects. If so, she will come to know not only how each name is pronounced in the two dialects, but also how some new name, previously unfamiliar to her, should be pronounced in the two dialects. In this case, there is difference in pronunciation of a linguistically significant sort, but the speaker will not consider the distinctness criterial; the difference in pronunciation will be considered to be within dialectal variation. Of course, the speaker might not be aware of the dialectal variation just referred to: the speaker might then take 'New York' and 'Noo Yawk' to be different words. We shall return to this matter below. For now, we may draw the conclusion that phonological distinctness, when it is believed to fall outside the range of dialectal variation, may be individuative of words. If it can be assumed that expressions that contain distinct words are occurrences of distinct expressions, then phonological distinctness may be individuative of syntactic expressions as well.

But what of the converse proposition: Does phonological identity underwrite expression identity? If two expression tokens are phonologically identical, does it follow that they are tokens of the same expression type? Certainly, if the tokens are categorially distinct (for example, the noun 'Ford' and the verb 'ford') the answer will be no, but what if they are categorially the same? One response would be to claim that there are ambiguous or indeterminate expression types (for example, 'Max'), and that context serves to fix their reference, on any given occasion of use. On that way of viewing things, there is, and can only be, one syntactic *expression* type 'Max', the value of its tokens being determined in a manner analogous to the valuation of indexicals. This approach presumes that the individuation of syntactic expressions is not itself fine-grained enough to distinguish noncoreferent occurrences whose phonological spelling is the same. But there is no good reason to believe this, and there is an alternative, which, it seems to us, is ultimately more compelling. First, consider the name 'Max', understood to be defined solely with respect to its phonological spelling. There are, of course, many people named 'Max', and it is natural to say that these people have the same name. It is also natural to say that some speakers but not others have the name 'Max' in their lexicons, and when they do, they have only one entry of this name, phonological spelling being unique. For these reasons, it makes sense to say that there is one *name* type 'Max'. But tokens of the name 'Max' may be deployed in many distinct syntactic expressions, themselves tokens of distinct *expression* types, and, when they are, those syntactic expressions refer to distinct individuals named 'Max'. On this way of looking at things, there is no value associated with the name 'Max' qua lexical item. In fact, on this account, names do not refer at all; reference accrues to syntactic expressions. While we commonly say that names refer (or are used to refer), this usage would be understood as informal.

There are two major assumptions behind the story just told. One is the distinction between names and the syntactic expressions that contain them. The other is that natural-language syntax is *formal*. We take it that for an account of language, including natural language, to be formal, it must be possible to determine, on examination of occurrences of its symbols, whether they are occurrences of the same symbol. No account not satisfying this requirement can be considered formal, although different kinds of accounts can impose the formality requirement in different ways. On the story just told, the symbols with respect to which natural languages are formal are the syntactic expressions, not just the names they may contain. This is the moral we draw from the problem of names with multiple bearers: if natural language is to be completely susceptible to a formal account, something other than sameness or difference of phonological spelling must be considered.

At this point, we need a notation which appropriately reflects this formality of natural language, one which precisely displays in syntactic structures the array of occurrences of expressions. We favor the indexical notation developed in our *Indices and Identity*,[3] which provides the minimal characterization needed: if NPs are occurrences of the same expression, they are coindexed; if they are occurrences of different expressions, they are not coindexed. Thus, there may be any number of occurrences of the expression type '$[_{\mathrm{NP}1}$ Max]' in a discourse (set of syntactic structures), all of which, if they refer at all, must refer to the same individual. Occurrences of an expression type corefer if they refer at all; all occurrences in a given discourse will be coindexed, and hence coreferential as a matter of representation. Each occurrence of '$[_{\mathrm{NP}1}$ Max]', however, will be an occurrence of a different expression type from any occurrence of '$[_{\mathrm{NP}2}$ Max]', and occurrences of different expression types may or may not refer to the same individual. This is not determined representationally; it is left open whether noncoindexed expression occurrences corefer or not.[4]

---

[3] Cambridge: MIT, 1994.

As we have observed, two expression occurrences may be spelled differently and yet be considered by a speaker to be occurrences of the same expression type. This was the case of dialectal variation. We can also have it that two expression occurrences may be spelled alike and yet regarded as occurrences of distinct expression types. This was the case of names with multiple bearers. A certain "autonomy" of syntactic expression is derived from such a view: identity of spelling appears to be detached from identity of syntactic expression. But can we also have it that there can be two distinct syntactic expressions that contain the same name and also have the same value? Yes, according to the condition we have given—although, it will turn out, no speaker would believe this to be so. This seeming paradox, we believe, is at the heart of Saul Kripke's referential puzzles, which we shall discuss in more detail below.

II. BELIEFS CONCERNING LANGUAGE

On this way of viewing syntactic expressions, the circumstances under which a speaker will believe that two expressions are translations of each other will be various. There is the case of distinct expressions that refer to the same thing and the case of occurrences of the same expression that are pronounced differently. The question is to determine the conditions under which one expression may be substituted for another. The notion of translation we wish to consider is a broad one; it embraces not only translation in the familiar sense, but anaphora as well. There is, therefore, no implication that translation is between distinct languages.

II.1. Translation. Let us call a statement of the form (1), which expresses a symmetrical relation between expressions, a *translation statement*:

(1) ⌜'*X*' translates '*Y*'⌝

where '*X*' and '*Y*' are schematic letters for expressions. Now, the term 'translation' has technical and nontechnical uses, and criteria for successful translation depend on what properties of the expressions to be related the translator is anxious to preserve. Those criteria, of course, are various. On our usage of the term, a translation statement is about linguistic equivalences, the sorts of linguistic equivalences of which the ordinary person is aware or about which she has opinions. Roughly put, a person will believe that two expressions are translations of each other, in our sense, if the person believes that the two expressions are covalued. The notion is actually not quite that simple, and we shall have occasion to return to this topic below.

---

[4] Notationally, what we have is not unlike what is done in logic, where, for instance, the variables which occur in formulas are distinguished as 'x_1', 'x_2', 'x_3', and so on. Any number of each variable may occur in a given proof, and while all occurrences of a variable will have a uniform assignment, occurrences of different variables may or may not be assigned the same values. Comparably to logic, where we would not normally have a different proof under a uniform permutation of the variables, we need an invariance principle for discourse so that they do not differ when different numerals are employed to display the same pattern of expression occurrences.

Translation statements express relations between two expressions (not lexical items), and those expressions may be coindexed or not. Thus, for indexed NPs, translation statements may be of the form (2) or (3):

(2) 'NP_i' translates 'NP_i'
(3) 'NP_i' translates 'NP_j'

The two forms have very different epistemic status. Grounds for believing translation statements of the form (2) are quite different from the grounds for believing translation statements of the form (3). This becomes clear when one considers cases.

Consider (2) first. We observed above that a speaker will believe that expression occurrences are occurrences of the same expression types only if she believes that any difference in their phonological spelling is due to linguistic variation. But there is a systematic exception to this found in every language, designated elements for which pronunciation is not criterial in the individuation of syntactic expressions. These elements are the pronouns. Two NPs may be occurrences of the same syntactic expression even though one may contain the name 'John' and the other the pronoun 'he'; binding theory proceeds on this assumption. Thus, we have translation statements such as (4):[5]

(4) '$[_1$John]' translates '$[_1$ he]'

The grounds for believing a translation statement such as (4) are purely linguistic in nature. Coindexed NPs are covalued, and, if covaluation is a sufficient license for translation, knowledge of the truth of (4) is grammatical in origin. The point is general: if two NPs are coindexed, they are translations of each other. Uniform substitution of another index in (4) will also yield a translation.[6] In that sense, translation of coindexed expressions holds irrespective of the values of the expressions. Given knowledge of various properties associated with an expression, and knowledge of English morphology, one knows the pronominal form of that expression.

---

[5] In our *Indices and Identity*, we constructed a theory of anaphora based on this conception, reserving the term 'vehicle change' for this relation. Included among vehicle changes are point-of-view translations: for example, translation between 'I' and 'you' or 'he', with indices kept constant. This has important consequences for invariance of logical form; full discussion lies beyond the scope of present concerns, however

Consider now a slightly different case. Some speakers believe that (5) is true; regional phonetic differences being for them no barrier to identity of expressions:

(5) '[$_1$New York]' translates '[$_1$Noo Yawk]'

Anyone conversant with the regional dialects of English would know that these are two different dialectal pronunciations of one English expression, and if they are conversant in this way, they will believe (5). But suppose someone does not know this. They might hear occurrences of 'New York' and occurrences of 'Noo Yawk' and not realize that they are different pronunciations of the same expression. Depending on circumstances, they might conclude that 'New York' and 'Noo Yawk' are names of different cities, or they might conclude that they are distinct names of the same city, no more instances of the same expression than 'Cicero' and 'Tully' are. If they come to the former conclusion, they will not believe (5); indeed, they will not believe that 'New York' and 'Noo Yawk' are translations of each other at all. If they conclude the second, they will believe (6):

(6) '[$_1$New York]' translates '[$_2$Noo Yawk]'

Notice that the grounds for believing (6) would be quite different from the grounds for believing (5). A speaker might come to believe (6) on the basis of hearing the same city called by two names; the speaker lacks the linguistic knowledge that the two names are dialectal variants of the same expression, and assumes that they are distinct expressions.

But now consider (7) and (8):

(7) '[$_1$London]' translates '[$_1$Rondon]'
(8) '[$_1$London]' translates '[$_1$Londres]'

If a speaker knows that 'London' and 'Rondon' are the English and Japanese pronunciations of the same expression, the speaker will believe (7); similarly for the French example. But the speaker might not know this; she might believe that the pairs are translations, but translations of different expressions. In that case, she would believe (9) and (10):

---

[6] Such substitution may yield an equivalent translation statement, if the statements are relative to equivalent discourses (roughly, discourses which differ only in the numerical values of the indices employed); otherwise, the translation statements will be of distinct expressions.

(9) '[$_1$London]' translates '[$_2$Rondon]'
(10) '[$_1$London]' translates '[$_2$Londres]'

Of course, the speaker might believe that 'Rondon' and 'Londres' are names of cities that are distinct from London and not translations at all.

The interlanguage circumstance just considered is in no way different from the dialectal one, and that is how it should be. There are, it should be mentioned, no linguistic criteria that define the boundaries between what we call dialects of the same language and what we call different languages. These sorts of distinctions are sociopolitical not linguistic in nature. More to the point, speakers do not appear to respect those boundaries either; they have beliefs about the individuation of expressions that transcend the distinction between difference of dialect and difference of language, however that distinction is made.

To summarize, we have the following inventory. Translation statements may hold between coindexed expressions. There are two cases of this kind to be distinguished. The "pronominalization" example constitutes a belief licensed by the machinery of grammar itself: the belief is analytic relative to grammar. The cognate examples constitute beliefs speakers have concerning the phonetic variation a single expression may display. Translation statements may also hold between noncoindexed expressions, and these are learned case by case.

In many instances, substitution salva veritate holds between expressions that are translations of each other. As we have seen, when translation relates coindexed expressions, this follows relatively trivially from knowledge of grammar. But when translation statements relate distinct expressions, matters become much more complex. So, suppose that we have (11) (setting aside the remote possibility that a speaker might somehow take 'Cicero' and 'Tully' to be dialectal phonetic variants):

(11) '[$_1$Cicero]' translates '[$_2$Tully]'

Given (11), we are licensed to move from (12) to (13):

(12) [$_1$Cicero] is [$_2$Tully]

(13) [$_1$Cicero] is [$_1$Cicero]

Since these are both true in all and only the worlds in which Cicero exists, the substitution is unproblematic. There is also no problem associated with moving by substitution from (14) to (15); provided that the final NPs in both are understood in the same way, both attribute to John a (trivial) belief of the self-identity of Cicero:

(14) John believes that Cicero is Tully.
(15) John believes that Cicero is Cicero.

Yet the move from (12) to (13) and (14) to (15) is not so simple; there is Frege's intuition that (12), but not (13) can represent an extension of knowledge and that (15) might be true and (14) false. Once we have developed a notion of "assignment," we shall be in a position to consider this problem.

II.2. Assignments. By asserting 'Cicero was a Roman', a speaker commits himself to certain beliefs. In particular, if the speaker is to speak sincerely, the speaker normally commits himself to the belief that Cicero was a Roman; if the speaker did not believe this, the speaker would not have spoken sincerely. It is immaterial to the sincerity of the speaker's beliefs that it might turn out, through further research into classical antiquity, that Cicero actually was not Roman at all, but rather Greek. The speaker is committed to speaking what he *believes* to be the truth; a speaker may be sincere, but incorrect. Another belief to which the speaker of 'Cicero was a Roman' commits himself is the belief that 'Cicero' refers to Cicero. The speaker might not be able to pick out Cicero by any definite description, and might, in fact, not be able to distinguish Cicero from Caesar by any means at his disposal. But for the sentence to be said sincerely, the speaker must believe that 'Cicero' refers to Cicero, whoever he might take Cicero to be.

Some of these beliefs might go wrong. The speaker might have stood in the Roman Senate, and, indicating Cataline, asserted 'Cicero is a great orator'. By saying that, the speaker may, through charity on the part of the hearer, be taken to mean that Cataline is a great orator, but confused as to Cataline's name; or in different circumstances, be taken to mean that Cicero is a great orator, but confused as to the identity of Cicero. In either case, the sincere speaker believes that he demonstrated the appropriate person and that 'Cicero' refers to Cicero.

Let us now focus on attributing the belief that Cicero was a Roman. If the speaker believes that Crassus believed that Cicero was a Roman, the speaker may say 'Crassus believed that Cicero was a Roman'. For this sentence to be true, it need not be the case that Crassus believed that 'Cicero' refers to Cicero. In fact, he may have never encountered the name 'Cicero', and, if asked, he might have denied that he had ever before had the belief that 'Cicero' refers to Cicero. Nevertheless, putting aside a kind of charitable attribution we shall consider momentarily, for the sentence to be true, it must be Cicero whom Crassus believed to be a Roman. To see this, consider the circumstance in which the speaker believes that Crassus knew the name 'Cicero' but that Crassus was confused about the identity of Cicero himself. He mistook Caesar for Cicero; when he encountered Caesar, he took him to be Cicero, and he never encountered Cicero himself. Suppose further that the speaker (and the hearer) are not confused in this regard. It seems clear that the speaker cannot say 'Crassus believed that Cicero was a Roman' to attribute the belief that Cicero was a Roman in this circumstance, at least not, as was mentioned, without assuming a certain charitable interpretation. Nor can he say 'Crassus believed that Cicero was a Roman' to attribute the belief that Caesar was a Roman. Although Crassus believed that 'Cicero' refers to Cicero, and would have said as much, he named the wrong man 'Cicero'; again, for the attribution to be correct, it must be Cicero whom Crassus believed to be a Roman.

In the cases just considered, it was assumed that the speaker was intending to be taken literally, and on that assumption, it was correct to say that, knowing Crassus's confusion, the speaker cannot attribute the belief to him that Cicero was a Roman. The hearer would naturally take the attributed belief to be about Cicero, not Caesar, and the wrong belief would be attributed. To change the example slightly, if Crassus took Plato for Cicero, and the speaker asserted that Crassus believed Cicero was a Greek, the hearer would naturally conclude that Crassus was mistaken as to the nationality of Cicero, not that Crassus had misidentified the unmentioned Plato as Cicero. Again, the speaker would have failed to attribute the correct belief to Crassus. But suppose that both the speaker and the hearer are aware, and the speaker believes that the hearer is aware, that Crassus took Plato for Cicero. Then it would seem that the speaker can say that Crassus believed Cicero was a Greek. An arch form of charity is at play here; the speaker is using the term 'Cicero' as he believes the agent uses it (incorrectly, as it happens) and, since he believes that the hearer is willing to go along with this, the correct belief can be attributed. We would emphasize that this example must be distinguished from the cases just considered, since here it is a requirement that the speaker believe that the hearer is aware of the agent's tions of the assertion itself, in which case they give rise to apparent exceptions to the assignment principle, to which we now turn.

III. BELIEF ASCRIPTION AND ATTRIBUTED ASSIGNMENTS

As mentioned above, if a speaker believes that expressions are translations of each other, it will be immaterial, at least sometimes, which expression is used. In that case, the speaker believes that the assignments associated with the expressions are *equivalent*, in that they assign the same value to different expressions.[7] But there are circumstances under which it is material which of a set of equivalent expressions is used. The speaker, for example, may wish to speak of the terms under which an agent holds a belief. In that case, the speaker may specify the particular assignment to which he believes the agent would assent. The speaker may, that is to say, attribute assignments. Let us consider the beliefs of speakers and the ways they may be reported in more detail.

The first thing to establish is that it is possible to attribute assignments. This occurs in example (17):

(17) John believes 'Cicero' has the value Cicero.

Here the attribution of an assignment is overt. But, in our view, there are also tacit occurrences. Consider (18):

(18) John believes Cicero was a Roman.

In (18), the belief is overtly attributed to John that Cicero was a Roman. But there is a way of saying (18) in which what is attributed in (17) is also attributed. Furthermore, this extra attribution is optional, so there are actually two logical forms for this sentence, one that includes a tacit assignment, and one that does not:

(19a) John believes [[Cicero$_1$ was a Roman] and ['Cicero$_1$' has the value Cicero$_1$]]

(19b) John believes [Cicero$_1$ was a Roman]

While for either of these to be true, it must be the case that John believes Cicero was a Roman, for (19a) to be true, John must believe that Cicero was a Roman and referred to by the 'Cicero' expression. If he fails to believe that Cicero was either or both of these, then (19a) is false. Thus, (19a) is false if John believes that Cicero was a Roman, but not that he is referred to by the 'Cicero' expression, or if he does not believe that Cicero was a Roman, but does believe that he is referred to by the 'Cicero' expression, or if he believes neither. There are two points of syntax that should be emphasized. First, the presence of the syntactic indices in (19a) indicates that we have, in the attribution, two occurrences of the expression 'Cicero' (excluding the occurrence within quotation marks). This is crucial to deriving the proper truth conditions for (19a), since it fixes that these two occurrences of 'Cicero' be covalued: for (19a) to be true, John must believe that the 'Cicero' expression refers to a person who was a Roman. Second, in (19a), *one* belief is attributed, that one belief being expressed as the conjunction of two sentences. Thus, the belief attributed in (19a) is to be distinguished from the conjoined attribution of beliefs, which could be expressed by a structure such as that in (20):[8]

---

[7] A bit more precisely, we say that assignments are *equivalent* if and only if they pair expressions with the same value. It now follows that an assignment is equivalent with itself, as well as to covalued assignments to distinct expressions

(20) John believes [Cicero was a Roman] and John believes ['Cicero' has the value Cicero]

This distinction is crucial; the truth conditions associated with (20) allow that John might believe, quite independently, that Cicero was a Roman, and that Cicero is referred to by the 'Cicero' expression, never having considered the two beliefs together. But the truth conditions associated with the structure in (19a) excludes that possibility; in (19a), John is said to have a single belief, which, as it happens, is expressed as the conjunction of two sentences. For him to have *that* belief, it must be the case that he believe that one person was a Roman and referred to by the 'Cicero' expression. In what follows, structures of the form in (19a) will be our focus, never that in (20).[9]

---

[8] As it happens, English can be structurally misleading in this regard, in ways that French, for example, is not: the English complementizer '*that*' may often be deleted, so that, even if one has a sentence of the form 'John believes that S and S'', it might be of the structure 'John believes [*that* [S and S']]' or 'John believes [*that* [S]] and [*that* [S']]'. In the first, a *single* belief, expressed as a conjoined sentence, is attributed; in the second, *two* beliefs are attributed, this latter form being, in effect, a reduced form of (20).

[9] A logical form such as (19a), in the terms of our account, is not itself subject to further analysis, in that there is no new information to be gained by appending an attributed assignment to an attributed assignment. Suppose that we were to analyze a logical form containing an attributed assignment in this way. There are two possibilities, depending on which term the attribution applies to:

(ia) John believes 'Cicero' has the value Cicero and 'Cicero' has the value Cicero.
(ib) John believes 'Cicero' has the value Cicero and ' 'Cicero' ' has the value 'Cicero'.

In neither of these logical forms does the assignment attributed add anything to what is otherwise provided by the primary attribution. In the former, it is simply redundant. In the latter, the information expressed by the assignment is known trivially, in virtue of knowing the language and the quotation-mark convention. (19a), as it stands, is completely analyzed

It is natural to call attributions containing assignments de dicto; their logical forms explicitly contain the expressions with respect to which the beliefs are held. By choosing to utter (18) under the logical form (19a), which includes a tacit assignment, rather than (19b), which does not, the speaker commits himself to a claim about the terms under which the agent holds a belief, and provides information as to what the agent believes over and above that given by the overt, primary attribution. For the attribution to be correct, it must be the case that the agent would agree that Cicero was a Roman referred to by the 'Cicero' expression. Recall, however, that the assignment principle requires that the sincere speaker, when using a sentence containing an expression NP_i, must believe an assignment. When assignments are attributed, however, the domain of the principle changes. Since the assignment resides within the scope of the attribution, the assignment is attributed to the agent, by the speaker. Under this circumstance, the requirement is that a sincere speaker, by attributing to an agent belief of an assignment, must believe that the agent believes the assignment. It is the speaker who is *using* the expression 'Cicero', of course, but he is using it so as to attribute the 'Cicero' assignment to the agent. He is using it as he believes the agent would. But why would a speaker choose to use an expression as the agent would, especially if he believes that he and the agent use the expression to refer to the same thing? Certainly, the most familiar reason, and the one that has been traditionally most emphasized, is that there is some *other* attribution, containing some *other* equivalent assignment, that the agent does *not* believe. If that constitutes the reason for the inclusion, then it stands as an *implicature* of an attribution containing an assignment that there is some other incorrect attribution, an attribution containing an assignment that the agent does not believe. To choose to include an assignment will, in this case, implicate that to have chosen a distinct assignment would have been incorrect.

To take an example, for (19a), with the included expression being 'Cicero', the excluded expression can be 'Tully', but not, for instance, 'Caesar', at least given a speaker with the normal beliefs. The speaker must hold that the included and excluded expressions refer to same thing, that their assignments are equivalent. Under the circumstance envisaged, this means that there is an implicature that the agent does not believe another attribution—say one containing a 'Tully' assignment—equivalent to the 'Cicero' assignment. Now, if an assignment is included in a belief attribution, it must be the assignment of an expression that is included in the overt belief attributed. This is for the usual conditions of relevance. Given that the relevancy requirement is just as much in force with respect to the excluded assignments as with respect to the included, what the speaker may implicate relative to (19a) is that an attribution such as that in (21), which is a logical form of 'John believes Tully was a Roman', would be incorrect:

(21) John believes [[Tully was a Roman] and ['Tully' has the value Tully]]

Observe that neither (19a) nor (21) entails the other. (19a) involves reference to the 'Cicero' expression and (21) involves reference to the 'Tully' expression. As these are distinct expressions, one cannot move by substitution from (19a) to (21), or vice versa; substitution within the quotes is prohibited. Of course, if one adds the premise that John believes that 'Cicero' translates 'Tully' and the premise that he has brought his 'Cicero' belief and his 'Tully' belief together, and the premise that John is logical, and perhaps other premises as well, the entailment from one to the other would go through. But there is no direct route from (19a) to (21) just by substitution. In any event, if the speaker believes that John does not believe that 'Cicero' translates 'Tully', which could be either because he believes that 'Cicero' and 'Tully' are expressions that refer to distinct individuals or because he does not know the expression 'Tully', he may correctly say (18) with the de dicto logical form (19a).[10]

But this is not the only reason a speaker might have for attributing an assignment. Consider again the case of the confused Crassus, described above. The point there was that, if the speaker believes that the hearer is aware of Crassus's confusion (that, as we would say, 'Cicero' is a name of Plato), he may say that Crassus believes that Cicero was a Greek. By a form of charity on the part the speaker, he may attribute the belief that Plato was a Greek, and by a form of charity on the part of the hearer he will be so understood. But in what does that charity consist? On the part of the speaker, it consists in using the expression 'Cicero' as he believes that the agent does, and on the part of the hearer it consists in taking it that way. The charity is arch, in that it does not correct a mistake but preserves it. Nevertheless, the point is that the attribution contains an assignment distinct from that which the speaker believes. Thus, this unusual case provides another reason for attributing an assignment.

---

[10] One might object on factual grounds that, given our requirement that, in the de dicto case, the agent must hold her belief under a particular assignment, our analysis of substitution failure is insufficiently general. There is the intuition that it is true to say of Lois that she believes that Superman flies, but not that she believes that Clark Kent flies even if she has never heard the names 'Superman' and 'Clark Kent'. Nothing that we have said would prevent this substitution But as Jennifer Saul pointed out—in "Substitution and Simple Sentences," *Analysis*, LVII, 2 (1997): 102-08—there is also the intuition that intersubstitution of these names is problematic even in what would seem to be transparent contexts, for example, 'Superman has more luck with women than Clark Kent'. This suggests that, whatever one's analysis of these "dual identity" cases, it may well be correct to claim that substitution failure derives from more than one source and that there is no general theory of substitution failure. De dicto attribution would constitute one limitation on substitution; whatever the principles are behind the dual identity cases would constitute another. We wish to thank Schiffer for first bringing these considerations to our attention.

One thing we are suggesting by our account is that there is a fundamentally *contrastive* element to the classical instances of de dicto belief ascription. There are, however, limiting cases. Imagine a religion in which 'God' is the one and only name allowed for the deity; and imagine that one only comes to know that name through the process of initiation. By assumption, there is no other assignment equivalent to the 'God' assignment which a person might know. Yet one might say of Max that he knows that God is great and attribute the 'God' assignment to Max, meaning to entail that he has acquired the 'God' assignment and thus that he has been initiated. If one wished to speak of contrast here, the contrast would be between believing an assignment and not believing one. But it might also be said that in this circumstance the normal contrastive implicature is suspended: if the speaker believes that the hearer believes that the deity has only one name, the implicature must be different in nature. We allow that there may well be other circumstances in which the attribution of assignments would be warranted. Nevertheless, the contrastive nature of the most interesting cases of de dicto belief is, we believe, correctly captured.

We have discussed (19a), the de dicto logical form of (18), at some length but, as we mentioned, there is also (19b), in which no assignment is attributed to the agent. Since no assignment is attributed, it is an implicature of (19b) that it is not relevant under what terms John holds the belief that Cicero was a Roman. The speaker may choose any assignment which is equivalent to: ' 'Cicero' has the value Cicero', without committing himself to any claim about which among equivalent assignments the agent might or might not hold; in this regard, (19b) is a non-de-dicto attribution. Understand non-de-dicto, to say (18) is the same as to say (22):

(22) John believes Tully was a Roman.

The irrelevance of the terms of attribution extends to cases in which the agent is a monolingual speaker of a language other than English.

(19b) is true just in case 'Cicero was a Roman' is a translation of a sentence in John's language that John believes.

As we use the terms, de dicto and non-de-dicto name types of attributions; both sorts of attribution are "semantically innocent." The distinction between them depends in part upon beliefs about whether or not expressions are translations of one another. But whether an agent believes that two expressions are translations of each other is independent of whether he believes that the reference of the expressions exists. Thus, the ambiguity that we find with (18) is just as much found with (23), which contains a so-called "empty name":

(23) John believes Santa Claus is an elf.

(23) will have two logical forms, (24a) and (24b):

(24a) John believes [[[$_1$Santa Claus] is an elf] and ['[$_1$Santa Claus]' has the value Santa Claus$_1$]]

(24b) John believes [[$_1$Santa Claus] is an elf]

The first will be said when, for example, the speaker wishes to have it as an implicature that there is some other attribution containing an assignment mentioning a name unfamiliar to John ('Kris Kringle', for example), which John does not believe. The second will be said when the terms of John's belief are not relevant. The difference between these logical forms will reveal itself, as before, when translation possibilities are considered. Suppose the speaker believes the translation statement (25):

(25) '[$_1$Santa Claus]' translates '[$_2$Kris Kringle]'

(24b) may then be translated as (26):

(26) John believes [[$_2$Kris Kringle] is an elf]

But (24a) cannot be translated in this way. One cannot, in particular, get to (27), since substitution within the quotes is prohibited:

(27) John believes [[[$_2$Kris Kringle] is an elf] and ['[$_2$Kris Kringle]' has the value Kris Kringle$_2$]]

What we see, then, is that sentences such as (23) are ambiguous between de dicto and non-de-dicto readings, and in exactly the same way that sentences that do not contain empty names are, admitting of the same possibilities of translation and attribution. This holds regardless of one's views about the reference of empty names; the distinction between de dicto and non-de-dicto attributions is independent of this.[11]

III.1. Translation of attributed assignments: the translated de dicto puzzle. It should be understood that attributed assignments can as much be "translated" as other attributed beliefs. Consider Sextus, a new member of the Roman Senate. He had heard the name '*Cicero*' (pronounced, in the nominative case, 'Kikero'), and the name '*Tullius*' (again the nominative case of the name rendered now in English as 'Tully'), but he was not aware that Cicero was Tully (he was not aware that, as he would have said, '*Cicero Tullius est*').[12] Suppose Sextus believed that Cicero was a Roman, but did not believe that Tully was. The following would be the logical form of the relevant ascription, as we would say it now in English:

---

[11] It might be argued that rather than analyze de dicto attribution as containing both a use and a mention of an expression, we could gain essentially the same results by allowing instead that a single expression plays dual semantic roles. For a suggestion along these lines, see Brian Loar, "Reference and Propositional Attitudes," *Philosophical Review*, LXXXI, 2 (1972): 43-62. The argument derives from the well-known example introduced by W. V. Quine in "Reference and Modality," in his *From a Logical Point of View* (Cambridge: Harvard, 1953), pp. 139-59, here p. 139:

(i) Giorgione was so-called because of his size.

On Quine's view, substitution failure of 'Barbarelli' for 'Giorgione' in giving (ii) from (i) shows that the occurrence of 'Giorgione' makes a dual contribution: Giorgione himself and the expression 'Giorgione'.

(ii) Barbarelli was so-called because of his size.

If the occurrence of 'Barbarelli' also makes a dual contribution, Barbarelli (= Giorgione), and the expression 'Barbarelli', the falsehood of (ii) is then explained. There is reason to doubt, however, that this argument supports the dual-role conclusion, since it neglects the role played by 'so'. Syntactically, 'so' is a pronoun, which as it occurs in (i) is incorporated into the verb; cf. (iii) from which it derives, which suffices to exclude (iv)

(iii) Giorgione was called so because of his size.
(iv) *Giorgione was so-called 'Giorgione' because of his size/*Giorgione was so-called so because of his size.

This suggests that (i) is syntactically parallel to (v), with 'so' taking as its value a name of the person referred to by the subject of the sentence:

(v) Giorgione was called 'Giorgione' because of his size.

Substitution cannot carry us from (i) to (ii) simply because the antecedent of 'so' is thereby changed; the result would be comparable to 'Barbarelli was called 'Barbarelli' because of his size'. Substitution is possible, however, if the antecedent of 'so' is not the subject, for instance, if it is in another sentence. If so, it is immaterial whether 'Barbarelli' or 'Giorgione' is the subject:

(vi) I was surprised to learn that Barbarelli was the Venetian painter Giorgione.
I was not surprised that Barbarelli/Giorgione was so-called because of his size

The role of anaphora in substitution is further shown by the true examples (vii) and (viii):

(vii) Giorgione was named for his size.
(viii) Giorgione was so-named for his size

Unlike with 'called', 'so' is optional with 'named', and substitution in (vii) yields the true (ix), while comparable substitution in (viii) yields (x), which is false with 'Barbarelli' as the antecedent of 'so'.

(ix) Barbarelli was named for his size.

(28) Sextus believed Cicero was a Roman and 'Cicero' had the value Cicero.

It is clear that Sextus had no contact with the English sentence ' 'Cicero' has the value Cicero'; but, of course, he did have contact with its Latin translation: ' '*Cicero*' *Ciceronis nomen est*'. It is a condition of the puzzle, as we have presented it, that the conjoined Latin sentence attributed to Sextus be translatable as a conjoined English sentence attributed to Sextus. These are presented in (29) and (30):

(29) Sextus believed that Cicero was a Roman and 'Cicero' had the value Cicero.
(30) Sextus believed that *Cicero Romanus erat et* '*Cicero*' *Ciceronis nomen erat.*

Consider the second clauses of the pair. For translation to go through, it will be required that ' '*Cicero*' ' and ' 'Cicero' ' (the names of the expressions '*Cicero*' and 'Cicero') be translations of each other. Similarly, it will be required that ' '*Tullius*' ' and ' 'Tully' ' (the names of the expressions '*Tullius*' and 'Tully') be translations of each other; these will appear in the attributed assignments that Sextus, by implicature, is said not to believe. But the de dicto translation relation we require must have it that ' '*Tullius*' ' (the name of the expression '*Tullius*') and ' 'Cicero' ' (the name of the expression 'Cicero') *not* be translations of each other.

Let us take it that translation, in this case, requires identity of value. It will be required, then, that '*Cicero*' and 'Cicero' be believed to be

---

(x) Barbarelli was so-named for his size.

On the dual-role account, it would have to be a special property of the predicate 'was so-called because of his size' (or perhaps just 'so-called') that requires that its subject make a dual contribution. Certainly, not all predicates are constrained in this way. What property would that be? The minimal contrast between (ix) and (x) tells the tale: it is the presence of the 'so' that makes the difference. If we are correct, then there is no expression that makes a dual contribution; it is 'so' that provides the second contribution.

[12] Since the English/Latin distinction is central to our point in this section, we shall use italics when citing Latin examples; English will continue to appear in Roman type.

occurrences of the same expression, just as it is required that *Cicero* and Cicero be believed to be the same person. We are not imagining that there are two ***kinds*** of translation here; there is just one kind of translation, proceeding at two different levels. ' 'Cicero' ' has as its value the expression 'Cicero' and ' '*Cicero*' ' has as its value the expression '*Cicero*', whereas 'Cicero' has as its value Cicero and '*Cicero*' has as its value *Cicero*. In the first instance, the translation relation will require that the expressions 'Cicero' and '*Cicero*' be believed to be identical, while in the second case the translation relation will require that Cicero and *Cicero* be believed to be identical.

In what sense may the expressions 'Cicero' in English and '*Cicero*' in Latin be believed to be identical? We argued above that speakers may believe varying pronunciations of the same name not to be criterial in distinguishing expressions. This is such a case. On this way of viewing things, there is one 'Cicero' expression, which, through its lifetime, has changed phonetic shape. Just as a person retains the same name while changing physical constitution from infancy to adulthood, so, too, does a word retain its identity as an expression while changing its phonetic constitution. (The point can be found in David Kaplan[15]). All of this being true, the translated de dicto attribution goes through, assuming, of course, that the speaker believes that 'Cicero' and '*Cicero*' are cognates.

To summarize, our claim has been that when a speaker attributes a belief, she may or may not commit herself to the terms under which the agent holds the belief. Thus, for us, the primary distinction to be drawn among belief ascriptions is between de dicto and non-de-dicto belief ascriptions. This distinction is linguistically represented by the presence or absence of attributed assignments. De dicto belief ascriptions are represented in logical forms involving (i) the overt belief ascribed, and (ii) attributed assignments to one or more of the syntactic expressions, dicta, out of which the overt ascription is formed. On the other hand, non-de-dicto belief ascriptions will involve nothing other than the overt belief ascribed. In their logical forms, there will be no attributed assignments associated with NPs, and, in that circumstance, any equivalent translation may be substituted for them.

IV SINGULARITY

We have now considered two kinds of belief about language that speakers have. The first, governed by the assignment principle, has it that the sincere use of expressions requires that the speaker believe the assignments associated with them. In the normal course of conversation (that is, where all participants are presumed to be sincere), each speaker will believe that all others speak in accordance with this principle, in virtue of believing that his interlocutors are competent speakers of the language. The assignments each speaker believes may vary, of course, reflecting the variance between speakers in the names in their lexicons. The second sort of belief we have considered are beliefs that expressions are translations of each other. These, too, may vary from speaker to speaker; one speaker may believe of a pair of expressions that they are translations of each other and another may believe of the same pair of expressions that they are not. Moreover, speakers believe that other speakers have beliefs of this sort, and again speakers may vary on whether or not they believe someone believes that expressions are translations.

---

[15] "Words," *Proceedings of the Aristotelian Society, Supplementary Volume* LXIV (1990): 93-119.

There is one other belief, which all speakers share, which we must recognize in order to complete our understanding of how speakers may deploy expressions in their language. To gain an intuitive grasp of its content, consider once more the case of Cicero. It is clear that a speaker may believe a 'Cicero' assignment, and that he may also believe a 'Tully' assignment, and that he may or may not know that Cicero and Tully are one and the same. But under what conditions may a speaker believe the assignment: ' '$[_1$Cicero]' has the value Cicero$_1$', and the assignment: ' '$[_2$Cicero]' has the value Cicero$_2$'? The answer we wish to give is that a speaker may believe assignments for two distinct but cospelled expressions of names just in case he believes that the assignments are not covalued. That is, with respect to any discourse and context, if a speaker believes two 'Cicero' assignments, he believes that there are two people named 'Cicero', while if he believes that there is only one person named 'Cicero', he will believe only one 'Cicero' assignment. We shall call this principle *singularity*.

> *Singularity principle:* If cospelled expressions are covalued, they are coindexed.

It is understood that there is great speaker-to-speaker variability with respect to assignments believed, and, as has been emphasized, that a speaker believes an assignment is no guarantee that he correctly uses the expression the assignment evaluates, for he may mistake the identities of the people to whom he refers. Nevertheless, the assignments a speaker believes are not completely unstructured. They are subject, as a class, to singularity.

Assignments, when attributed in the context of de dicto attributions, must also accord with singularity, but in this case it is with respect to the agent's, not the speaker's, beliefs. To see the effect of this, we turn to a pair of puzzles well known since their introduction by Kripke (*op. cit.*). In the first of these, it is supposed that a speaker believes that John believes, for whatever reason, that there are two people named 'Paderewski', but that the speaker believes there is only one such person. Against this background, the speaker may say (31) and (32) without contradiction:[14]

(31) John believes that Paderewski is a genius.
(32) John does not believe that Paderewski is a genius.

Why are (31) and (32) not contradictory? Their logical forms will be as follows:

(33) John believes that Paderewski_1 is a genius and 'Paderewski_1' has the value Paderewski_1
(34) John does not believe that Paderewski_2 is a genius and 'Paderewski_2' has the value Paderewski_2

The logical forms of the *that* clauses in (31) and (32) are indexically distinguished in (33) and (34). Therefore, (31) and (32) say nothing inconsistent about John, since the beliefs attributed to John are distinguished. This is in accordance with the standard intuition concerning this puzzle, which has it that the speaker and agent are both perfectly logical. One of them, perhaps even both, are, of course, wrong about the facts, but this makes no difference. What matters for the puzzle is that the speaker's and agent's beliefs diverge as described. Since the assignments in (33) and (34) are attributed, only the agent must believe them, and since the agent believes them in accordance with singularity, it is he who is said to believe that there are two people named 'Paderewski'. It is not a consequence of (33) and (34) that the speaker holds this belief.

Since it is explicit in (33) and (34) that the agent believes two 'Paderewski' assignments, the de dicto nature of this puzzle, first emphasized by Kripke, is captured.[15] The logical forms (33) and (34), however, are not the only ones which (31) and (32) could have; they could also have logical forms with no attributed assignments. These logical forms would also be consistent, but would not comport with the conditions of the puzzle; for it would now follow from the assignment principle that the *speaker* believes that there are two 'Paderewski' assignments, and from singularity that the speaker believes that there are two people named 'Paderewski'. Of course, in a different scenario, the speaker could believe that there are two Paderewskis, and if he did, he could make the first-person ascriptions: 'I believe Paderewski$_1$ is a genius' and 'I do not believe Paderewski$_2$ is a genius'; given singularity, it follows from these self-ascriptions that he believes there are two people named 'Paderewski'. He may be right or wrong about this, but if there is just one Paderewski, then these first-person ascriptions are comparable to those in the third-person with respect to the puzzle scenario. But there is a difference between the cases. Since the speaker is the agent of the belief in the first-person ascriptions, there is no disparity in beliefs between speaker and agent of the sort that gives rise to the de dicto character of third-person ascriptions. Thus, in this case, the puzzle has a non-de-dicto nature, characterized by logical forms from which assignments are withheld.[16]

---

[14] In presenting this puzzle, the story is often embellished by saying, for instance, that the agent believes that there are two people: one a statesman, the other a pianist. Putting matters this way, however, makes it appear as if the agent's having distinct descriptions is somehow part of the puzzle. This is incorrect. All that is required to generate this version of the puzzle is that the agent believe there are two people with the same name in contrast to the beliefs of the speaker—that the agent believe two 'Paderewski' assignments.

[15] Taken de dicto, a speaker uttering (31) and (32) says that the agent has one belief, but does not have another, distinct belief; when (31) is paired with 'John

Singularity specifies of cospelled expressions that, if they are covalued, they are occurrences of the same expression. We have just contrasted the effects of singularity as it applies to assignments attributed to an agent with its effects when it applies to the assignments of the speaker. The contrast arises when the speaker and agent disagree as to the number of people named 'Paderewski'. The possible effects of variation between speaker and agent with respect to beliefs concerning translation are not addressed in this scenario. But translation will present similar kinds of problems, since different occurrences of an expression that are not cospelled may be translations of each other and covalued. As we pointed out in the translated de dicto puzzle, a speaker may believe that the English word 'Cicero'

---

believes that Paderewski is not a genius', the speaker says that the agent has two distinct beliefs. This follows for the latter case by the same analysis as for the former; that is, by their having logical forms in which assignments are attributed.

[16] Normally, first-person reports in the present tense will be non-de-dicto; they will only be de dicto to the extent that the speaker/agent wishes to contrast his own beliefs relative to assignments that he believes to be equivalent. It is perhaps difficult to imagine a circumstance in which, given a belief, he would want to say he has it relative to a 'Cicero' assignment, but not relative to a 'Tully' assignment, given, of course, that he believes that they are equivalent assignments. Otherwise, he would have no grounds for contrasting them, so there would be no need for him to commit himself to a belief with respect to a specified assignment.

and the Latin word '*Cicero*' are phonetic variants of the same expression, and thus, trivially, translations of each other. But what are the effects when speakers and agents have different translational beliefs? To answer this, we turn to the second of Kripke's puzzles.

To set the stage, Pierre, a monolingual French speaker, has read about London and has come to believe that London (or, as he would have said, 'Londres') is beautiful. He then moves to London and learns English. Becoming familiar with the city, he comes not to believe that London is beautiful. He has, of course, made no logical error, nor has he changed his mind; he is simply unaware that London is Londres. Now imagine a speaker, who believes that 'London' and 'Londres' are occurrences of the same expression, reporting Pierre's beliefs by uttering (35) and (36):

(35) Pierre believes that Londres is beautiful.
(36) Pierre does not believe that London is beautiful.

If 'London' and 'Londres' are taken as the *speaker* takes them, as occurrences of the same expression, then the speaker will have spoken inconsistently. He will not have spoken this way, however, if, in common with the "Paderewski" scenario, (35) and (36) have logical forms containing attributed assignments:

(37) Pierre believes that Londres$_1$ is beautiful and 'Londres$_1$' has the value Londres$_1$
(38) Pierre does not believe that London$_2$ is beautiful and 'London$_2$' has the value London$_2$

Recall that there are two things of which Pierre is unaware. One is that London is Londres, that they are the same place. The other is that 'London' and 'Londres' are phonetic variants of the same expression. If he believed the second, the first would follow trivially (although if he believed the first, he might not believe the second: he might know that London is Londres but not believe that 'London' and 'Londres' are cognate, and therefore variants of the same expression (cf. 'John' and 'Ian')). It is clear, then, that the expressions mentioned in the assignments in (37) and (38) cannot be coindexed, since, if they were, and Pierre brought all of his beliefs together, he could conclude just from them alone that London is Londres. But the conditions of the puzzle are such that he must not be able to conclude that just by dint of logic. Pierre would report that London and Londres are different cities, from which it can be inferred that he could not hold that 'London' and 'Londres' are variant expressions referring to the same city.

In the "Paderewski" scenario, given the beliefs of the speaker, singularity required that the speaker believe one 'Paderewski' assignment; and, given the beliefs of the agent, singularity required that the agent believe two 'Paderewski' assignments. In the present case, because 'London' and 'Londres' are not cospelled, singularity allows that the speaker believe both a 'London' assignment and a 'Londres' assignment, and that the speaker believe that 'London' and 'Londres' are not only translations of each other, but occurrences of the same expression. Of course, the speaker in the "Paderewski" scenario believes that 'Paderewski' and 'Paderewski' are not only translations of each other but occurrences of the same expression, since he believes that there is only one 'Paderewski' expression and every expression is a translation of itself. In this respect, the "Paderewski" and "Londres" scenarios are similar. Note, however, that from the fact that 'London' and 'Londres' are believed to be occurrences of distinct expressions it does not follow that they are believed to have different values; they might be held to be like 'Cicero' and 'Tully'. But from the fact that 'Paderewski$_1$' and 'Paderewski$_2$' are believed to be occurrences of distinct expressions it does follow, by singularity, that they are believed to have different values. Thus, in contrast to the "Paderewski" scenario, it does not *follow* from (37) and (38) that Pierre believes that there are two cities; all that follows is that Pierre does not believe that 'London' and 'Londres' are occurrences of the same expression (and thus translations). But, given the assumption that Pierre is logical, this is enough; for if Pierre believed that 'London' and 'Londres' were distinct but covalued expressions, he could hardly hold that Londres is beautiful and London not. What the reports of Pierre's beliefs about London and John's beliefs about Paderewski have in common, then, is that each pair contains equivalent assignments to distinct expressions. Where they diverge, however, is just at the point at which singularity comes into play.

V. IDENTITY PUZZLES

In essence, the assignment principle and singularity, as aspects of a speaker's linguistic competence, are conditions on *expressability* in language. They determine, in part, how a speaker may deploy expressions, given his beliefs about their reference. Such beliefs are directly reflected in the identities each sort of speaker would assert. Any speaker who has the name 'Paderewski' in his lexicon could assert (39):

(39) Paderewski$_1$ is Padereswki$_1$

but a speaker who believes there are two people with the name in question could assert something else consonant with his beliefs, the nonidentity in (40):

(40) Paderewski$_1$ is not Padereswki$_2$

The speaker who believes there is only one person named 'Paderewski' would, of course, not assert (40), because it entails something which, given the assignment principle and singularity, he believes to be false—that there are two people named 'Paderewski'. But that he would not assert this does not bar him from attributing it to someone else, provided that he does so de dicto. Thus, he may say (41), with the logical form in (42):

(41) John believes Paderewski is not Paderewski.

(42) John believes [Paderewski$_1$ is not Paderewski$_2$ and 'Paderewski$_1$' has the value Paderewski$_1$ and 'Paderewski$_2$' has the value Paderewski$_2$]

The puzzle as to how (41) can report a consistent belief thus echoes the puzzle that has been under discussion in the previous section.

In this case, as with all the puzzle cases considered, it is immaterial whether it is the speaker or hearer (or neither) who is factually correct. Rather, it is the conflict in their beliefs that is significant, and that conflict might go in different directions. So suppose that a speaker, who believes that there are *two* people named 'Paderewski' wishes to tell a hearer, who believes the same, that someone else believes they are one and the same. The speaker could accomplish this by just attributing the opposite of (41), that is, by saying (43):

(43) John believes Paderewski is Paderewski.

What are we to take the logical form of (43) to be? If the embedded sentence is indexed as in (39), then (43) would attribute a true but trivial belief to John. As we have described it, however, this is not what the speaker of (43) wishes to say. Rather, he wishes to say to the hearer that the agent, incorrectly in their view, identifies two people that they distinguish. That (43) may be used to say this follows from its having the following logical form:

(44) John believes Paderewski$_1$ is Paderewski$_2$

Examining (44), since there are no assignments attributed, we may conclude from the assignment principle and singularity that the 'Paderewski' expressions it contains are used in accordance with the speaker's belief that they are not covalued. Nothing, however, follows in virtue of their occurrence about the *agent's* beliefs about assignments. Rather, what the agent believes is *entailed* by the attribution. From (44), it follows that what is distinguished by the speaker by his use of distinct 'Paderewski' expressions is equated by the agent, and from this can be derived, given the dictates of the assignment principle and singularity, that, if the agent knows the name 'Paderewski', then he believes only one 'Paderewski' assignment, not two.

In thus reversing the beliefs of the speaker and the beliefs he wishes to attribute, we observe a shift from de dicto to non-de-dicto attribution. Indeed, in this context, if the attribution were to be de dicto, so that the embedded clause of the logical form of (43) would be:

(45) [Paderewski$_1$ is Paderewski$_2$], and 'Paderewski$_1$' has the value Paderewski$_1$ and 'Paderewski$_2$' has the value Paderewski$_2$

the inconsistency would be attributed to John that he believes that there are two people with the name 'Paderewski' (given the attributed assignments), and that he believes that only one person is named 'Paderewski' (given the truth of the primary attribution). We should note something about the logical form (45), however. Its inconsistency is not inherent; it only becomes so when placed in overtly attributed contexts. Outside such contexts it may be asserted. So suppose we have a speaker who, as in the initial scenario, believes there is just one person named 'Paderewski' and an addressee, John, who believes there are two, an addressee, that is, who would be prepared to assert (40). In that circumstance, to straighten the addressee out, the speaker may deny (40) by uttering (46):

(46) But John, Paderewski is Paderewski.

The speaker would not achieve this goal, however, if he were taken to have said 'Paderewski$_1$ is Paderewski$_2$', since this would only entail something about his own assignments, something in fact false, that he believes there are two people named 'Paderewski'. Rather, the logical form of (46) must be as in the attribution in (45), in which, since there are attributed assignments, it follows that it must be someone other than the speaker who holds this belief. The only possibility in this context is the addressee. Thus, by an utterance of (46) in denial of (40), the addressee is asked to believe that two of *his* assignments are equivalent. The speaker's assertion to the addressee is, in effect, *your* 'Paderewski' expressions corefer. Taken this way, an utterance of (46) would be cause for the addressee to alter his beliefs; it would act as a corrective to what are, from the speaker's point of view, the mistaken beliefs of the addressee.

It would appear from this case that de dicto interpretation of NPs is in no way restricted to their occurrence in *that* clauses. This observation is strengthened when we shift our attention to the puzzle of London and Londres, where to straighten Pierre out, a speaker may address him and say, 'Londres is London'. Analyzing this in a parallel fashion, the relevant logical form would then be:

(47) Londres$_1$ is London$_2$ and 'Londres$_1$' has the value Londres$_1$ and 'London$_2$' has the value London$_2$

The speaker in saying (47) is attributing assignments to the addressee Pierre, who is thereby invited to believe that the values named in his two assignments are the same.[17] The parallelism between this case and the 'Paderewski' case is not, however, complete. So, consider Pierre's situation. He does not know that Londres is London and he does not know that 'London' and 'Londres' are occurrences of the same expression. Consider the ignorance of John. He does not know that there is one Paderewski and he does not know that there is one 'Paderewski' expression. To relieve Pierre's ignorance, it is enough to say de dicto that London is Londres. To relieve John's ignorance, it is enough to say de dicto that Paderewski is Paderewski. In John's case, given singularity, he must revise his assignments; there will be one 'Paderewski' assignment instead of two. In Pierre's case, he must revise his assignments in such a way that the 'Londres' assignment and the 'London' assignment have the same value. But he need not conclude that the 'Londres' assignment and the 'London' assignment are assignments of coindexed expressions, since he must also become apprised of the fact that 'London' and 'Londres' are variants of the same expression. One could apprise him of this explicitly, of course, but, short of that, Pierre could hold, in the face of being told that London is Londres, that 'London' and 'Londres' are no more variants of the same expression than 'Cicero' and 'Tully' are. It is at this point that the treatment of the "Paderewski" and "Londres" puzzles diverge.

This divergence can be observed further, reverting to the reverse scenario introduced above, in which the hearer believes there is one person named 'Paderewski', the speaker, two. The speaker then tries to straighten the hearer out by uttering: 'Paderewski is not Paderewski'. It is clear that the speaker's utterance will not alone suffice to meet this goal; for consider the form of the speaker's utterance. It cannot contain attributed assignments, since the speaker believes that the addressee believes singularity and therefore will have only one 'Paderewski' assignment. Thus, the logical form of the speaker's utterance will be as in (48):

---

[17] There are two other logical forms, both non-de-dicto, which are in principle possible for 'Londres is London'. One is 'Londres$_1$ is London$_1$'. While for the speaker, this is a translational triviality, since Pierre believes that Londres is not London, he cannot believe that 'Londres' and 'London' are variant expressions for the same city. So it is hard to see how this could be the logical form. The second is 'Londres$_1$ is London$_2$'. But the speaker would utter this only if he did not believe 'Londres' and 'London' are cognates, and then we would only have a variant of the standard 'Cicero'/'Tully' case

(48) Paderewski$_1$ is not Paderewski$_2$

How is the hearer to take (48)? He might reason as follows: If the speaker is not uttering a contradiction of the form (49) (and why would any speaker do that?), the speaker's sentence must have the form (48).

(49) Paderewski$_1$ is not Paderewski$_1$

But, if the sentence has the form (48), singularity, which the speaker must believe, dictates that the speaker believes that there are two 'Paderewski' expressions. The hearer could then conclude that the *speaker* believes that there are two people named 'Paderewski', but he would not be enlightened by an utterance of (48) about anything he himself holds about Paderewski. The speaker, unlike in the case just discussed, would not be denying anything that the hearer holds to be the case, and the hearer will be left in the dark by (48) as to what two people the speaker is distinguishing.

But what of Pierre, in a comparable reverse scenario, where we suppose he believes that there is one city with the two names 'London' and 'Londres,' and also believes that 'London' and 'Londres' are variants of the same expression. Pierre believes (50):

(50) '[$_1$London]' translates '[$_1$Londres]'

To inform Pierre that he is incorrect, the speaker may utter:

(51) London$_1$ is not Londres$_2$

But this utterance, in contrast to the formally identical (48), succeeds to the desired end. Why should this be?

By believing (50), Pierre believes that the expressions 'London' and 'Londres' are variants of the same expression, and that, since they are, they are covalued. But he does not hold that this belief is analytic. He knows that he might be wrong that the two names are translations. Thus, when the speaker says (51), he may bring together his translational belief and the assertion of the speaker and find them to be in conflict, and, if he decides to believe the speaker, he will accept (51) as corrective. Matters are quite different in the case of John. He believes that (52) is true:

(52) Paderewski$_1$ is Paderewski$_1$

but (52) really is analytic, and is not, therefore, a candidate for revision.

We may now raise an important point concerning (37) and (38), repeated here.

(37) Pierre believes that Londres$_1$ is beautiful and '[$_1$Londres]' has the value Londres$_1$

(38) Pierre does not believe that London$_2$ is beautiful and '[$_2$London]' has the value London$_2$

Supposing that the speaker believes the translation statement (50), he will consider (53) a translation of (37), since he considers 'London' and 'Londres' to be variants of the same expression:

(53) Pierre believes that London$_1$ is beautiful and '[$_1$London]' has the value London$_1$

Therefore, in principle, the speaker can correctly describe Pierre's problem by uttering (38) and (53). There would have to be some sort of preparation for this, however; since the distinction between the two 'London' expressions is inaudible, a hearer would not otherwise be able to distinguish them. Now, suppose someone (we have Kripke in mind) were to ask the question:

(54) Does Pierre, or does he not, believe that London is beautiful?

The hearer of (54), if he understands it as a de dicto attribution, would have to know which question is being asked; on one indexing of the question, the appropriate answer is (38), while on the other indexing, the appropriate answer is (53). On the other hand, if it is understood as a non-de-dicto attribution, then Kripke is assuredly right that the question cannot be answered as asked, since it is unambiguous, the speaker believing, unlike Pierre, that there is only one place London, and given the assignment principle and singularity, that there is only one expression 'London'.[18]

---

[18] On Kripke's view, all of the belief puzzle cases, including the standard cases of substitution failure of distinct but coreferring names (for example, 'Cicero' and 'Tully'), arise from what he sees as paradoxical properties of belief reporting. Crucially for him, this includes cases that he believes cannot be analyzed as arising from failures of substitution—for example, the "Paderewski"-puzzle, in which he maintains there is no distinction comparable to that between 'Cicero' and 'Tully'. Our disagreement with Kripke lies here; what we are saying is that Kripke's discussion is predicated on a misanalysis of the notion of linguistic expression. Once this is in place, the puzzle cases, understood de dicto, become uniform with the standard cases of substitution failure, but in a way, we note, compatible with a Millian theory of names. 'Cicero' can be substituted for 'Tully', for they corefer, but ' 'Cicero' ' cannot be substituted for ' 'Tully' ', as what they refer to are plainly not identical; they are two different expressions (leaving aside aberrant beliefs about translation). Similarly, 'Paderewski$_1$' can be substituted for 'Paderewski$_2$', but not ' 'Paderewski$_1$' ' for ' 'Paderewski$_2$' ', for they, too, refer to different expressions.

VI CONCLUDING REMARKS

Our goal in this paper has been to present a theory that incorporates the view that syntactic expressions of natural language play the role of modes of presentation. By attributing an assignment to an agent, a speaker says that the agent holds a belief with respect to a certain mode; but to attribute that belief under a different but equivalent assignment is to attribute with respect to a different mode, and is not to attribute the same belief. On our way of conceiving of "substitution," the central consideration thus becomes the individuation of syntactic expressions and the conditions that govern their deployment in speaker's utterances.

A speaker may have firm grounds for attributing semantic beliefs to an agent; in the case of de dicto ascription, a speaker may know, for example, that the agent uses one expression and not another to refer to an object. The speaker can know such things about an agent while knowing nothing about the qualities that the agent associates with the objects to which the agent refers. A person may come to the conclusion that there are two people with the same name, and then deploy two cospelled expressions to refer to those people. But what the grounds would be for his coming to the conclusion that there are two different people, or how he thinks of them, is a different matter altogether. However an agent may conceive of an individual, and whatever role this may play in how one associates a name with a value, such conceptions are not modes of presentation in our theory.[19]

The virtue of taking expressions as modes of presentation is that there are precise criteria by which the modes may be individuated.[20] The analogy of syntactic expressions to modes of presentation as classically conceived is not complete, however. As Frege had it, the modes of presentation were senses of expressions, but this created difficulties. In particular, he had to face the fact that the sense of an expression may vary widely from speaker to speaker. Given his theory of indirect reference, this had the consequence that it would be at best unlikely that one could ever attribute the same belief to two persons. Ultimately, Frege was forced to concede, in "The Thought,"[21] that "we must really stipulate that for every proper name there shall be just one associated manner of presentation of the object so designated" (*ibid.*, p. 12). That requirement played the role of singularity in Frege's theory. But he had no right to the requirement, which flew in the face of the facts. Of course, Frege had early on decided that expressions themselves could not serve the role of modes of presentation, the result being his misidentification of singularity as a condition on senses, not, as it should be, a condition on syntactic expressions. Our version of singularity is much more natural. In the case of shared belief, if speakers believe that there is a person named 'Cicero', then they believe that there is one 'Cicero' expression that refers to him. Furthermore, that 'Cicero' expression may serve as a mode of presentation of Cicero in a belief attributed to them. Admittedly, they might not put their belief in exactly these terms. They might say that Cicero has a name and that that name is 'Cicero'. But they will agree that Cicero has only one name 'Cicero'. That is the essence of singularity, as we conceive of it.

---

[19] Given these remarks, it should be clear that we would not liken indices to footnote markers, where the footnotes themselves describe how an agent conceives of an individual, which seems to be what is envisaged in Akeel Bilgrami, *Belief and Meaning* (Cambridge: Blackwell, 1992).

[20] Thus meeting a challenge set by Schiffer, that words cannot play the role of modes of presentation "until we are given the intended method of word individuation" without "appeal to word meanings" (*op. cit.*, p. 257).

[21] In P.T. Geach, ed., *Logical Investigations* (Cambridge: Blackwell, 1977), pp. 1-30.

To conclude, then, the analysis we have presented seems to us to confirm the distinction between names and expressions we have drawn, as well as the utility of that distinction in explaining belief attribution. We have argued that speakers' beliefs concerning assignments, as well as their beliefs concerning the beliefs of others concerning assignments, may be syntactically represented in the logical forms of sentences. This lies at the heart of the de dicto/non-de-dicto distinction we have advocated. The role of syntax in these matters is thus greater, in our opinion, than has generally been recognized. As Kripke emphasized, there is nothing paradoxical in the various puzzles we have discussed; there are perfectly straightforward descriptions that accommodate all the facts. The problem, of course, is to discover the terms in which the description should be stated. The syntactically based description presented here, we have argued, does that well.

ROBERT FIENGO

Queens College and the Graduate Center,
City University of New York

ROBERT MAY

University of California/Irvine

WHAT IS SO BAD ABOUT CONTRADICTIONS?*

I shall address the title question, and the answer I shall give is: maybe nothing much. Let me first explain how, exactly, the question is to be understood. I shall interpret it to mean: What is wrong with believing *some* contradictions? I emphasize the 'some'; the question 'What is wrong with believing *all* contradictions?' is quite different, and, I am sure, has a quite different answer. It would be irrational to believe that I am a fried egg. (Why, we might argue about, but that this is so is not contentious.) A fortiori, it is irrational to believe that I am both a fried egg and not a fried egg. It is important to emphasize this distinction right at the start, since the illicit slide between 'some' and 'all' is endemic in discussions of the question, as we shall see.

I think that there is nothing wrong with believing some contradictions. I believe, for example, that it is rational (rationally possible—indeed, rationally obligatory) to believe that the liar sentence is both true and false. I shall not argue for this directly here, though. I have discovered, in advocating views such as this, that audiences suppose them to be a priori unacceptable. When pressed as to why, they come up with a number of arguments. I shall consider five of the most important, and show their lack of substance. They can be summarized as follows:

(1) Contradictions entail everything.
(2) Contradictions cannot be true.
(3) Contradictions cannot be believed rationally.
(4) If contradictions were acceptable, people could never be rationally criticized.
(5) If contradictions were acceptable, no one could deny anything.

I am sure that there must be other possible objections, but the above are the most fundamental that I have encountered. I shall take them in that order. What I have to say about the first objection is the longest, because it lays the basis for all the others.

I. OBJECTION ONE. CONTRADICTIONS ENTAIL EVERYTHING

The objection is that rational belief is closed under entailment, but a contradiction entails everything. Hence, if someone believed a contradiction, they ought to believe everything, which is too much.

I certainly agree that believing everything is too much: I have already said that there is an important difference between *some* and *all*

---

* The essay is a written version of a lecture that has been given at universities in South Africa, Canada, and the United States in the last few years. I am grateful to the audiences for numerous lively discussions.

0022-362X/98/9508/410–26

here. Still, I take the argument to be unsound. For a start, it is not at all obvious that rational belief is closed under entailment. This seems to be the lesson of the "paradox of the preface." You write a (nonfictional) book on some topic—history, karate, cooking. You research it as thoroughly as possible. The evidence for the claims in your book, $\alpha_1,\dots,\alpha_n$, is as convincing as empirically possible. Hence, you endorse them—rationally. Nonetheless, as you are well aware, there is independent inductive evidence of a very strong kind that virtually all substantial factual books that have been written contain some false claims. Hence, you also believe $\neg(\alpha_1 \wedge \dots \wedge \alpha_n)$—rationally. But you do not believe $(\alpha_1 \wedge \dots \wedge \alpha_n) \wedge \neg(\alpha_1 \wedge \dots \wedge \alpha_n)$, a simple contradiction, even though this is a logical consequence of your beliefs. Rational belief is not, therefore, closed under logical consequence.

This is all just softening-up, though. The major problem with objection number one is the claim that contradictions entail everything: α, $\neg\alpha \vdash \beta$, for all α and β. The Latin tag for this is *ex contradictione quodlibet*. I prefer the more colorful: *explosion*. It is true that explosion is a valid principle of inference in standard twentieth-century accounts of validity, such as those of intuitionism and the inappropriately called classical logic. But this should be viewed in a historical perspective.

The earliest articulated formal logic was Aristotle's syllogistic. This was not explosive. To see this, merely consider the inference:

Some men are mortals.
No mortals are men.
Hence all men are men.

This is not a valid syllogism, though the premises are inconsistent. According to Aristotle,[1] some syllogisms with inconsistent premises are valid, some are not. Aristotle had a propositional as well as a syllogistic logic. It was never clearly articulated, and what it was is rather unclear. For what it is worth, however, this does not seem to have been explosive either. In particular, a contradiction, $\alpha \wedge \neg\alpha$, does not entail its conjuncts.

The Stoics did have an articulated propositional logic. But while one might try to extract explosion from some of the theses they endorsed, it is notable that it is not to be found in anything that survives from that period—and one would expect any principle as striking as this to have been made much of by the most notable critic of Stoicism, Sextus Empiricus. Presumably, then, explosion was not taken to be correct by the Stoics.

So, if explosion is not to be found in ancient logic, from where does it come? The earliest appearance of the principle of which I am aware seems to be in the twelfth-century Paris logician, William of Soissons. At any rate, William was one of a school of logicians called the "Parvipontinians," who were well known, not only for living by a small bridge, but also for defending explosion.[2] After this time, the principle appears to be a contentious one in medieval logic, accepted by some, such as Scotus; rejected by others, such as the fifteenth-century Cologne school.

---

[1] *Prior Analytics* 64^{a}15 Cf my "Negation as Cancellation, and Connexivism," *Topoi* (forthcoming).

The entrenchment of explosion is, in fact, a relatively modern phenomenon. In the second half of the nineteenth century, an account of negation—now often called Boolean negation—was championed by George Boole, Gottlob Frege, and others. Boolean negation is explosive, and was incorporated in the first contemporary formal logic. This logic, now usually called classical logic (how inappropriate this name is should now be evident), was so great an improvement on traditional logic that it soon became entrenched. Whether this is because it enshrined the natural light of pure reason, or because it was the first cab off the rank, I leave the reader to judge.

There is, in fact, nothing sacrosanct about Boolean negation. One can be reminded of this by the fact that intuitionists, who gave the second contemporary articulated formal logic, provide a different account of negation. Despite this, intuitionist logic is itself explosive. Logics in which explosion fails have come to be called *paraconsistent*. The modern construction of formal paraconsistent logics is more recent than anything I have mentioned so far. The idea appears to have occurred to a number of people, in very different countries, and independently, after World War II. There are now a number of approaches to paraconsistent logic, all with well-articulated proof theories and model theories.

I do not intend to go into details here. I shall just give a model-theoretic account of one propositional paraconsistent logic, so that those unfamiliar with the area may have some idea of how things might work.[3] I assume familiarity with the classical propositional calculus. Consider a language with propositional parameters, p, q, r,... and connectives $\wedge$ (conjunction), $\vee$ (disjunction), and $\neg$ (negation). In classical logic, an evaluation is a *function* that assigns each formula one of 1 (true) or 0 (false). Instead of this, we now take an evaluation to be a *relation*, R, between formulas and truth values. Thus, given any formula, α, an evaluation, R, may relate it to just 1, just 0, both, or neither. If $R(\alpha, 1)$, α may be thought of as true under R; if $R(\alpha, 0)$, it may be thought of as false. Hence formulas related to both 1 and 0 are both true and false, and formulas related to neither are neither true nor false.

---

[2] For references and more details of the following history of paraconsistency, see my "Paraconsistent Logic," in D. Gabbay and F. Guenthner, eds., *Handbook of Philosophical Logic* (Dordrecht: Kluwer, 2nd edition, forthcoming).

[3] The logic is that of first-degree entailment. For further details of all the approaches to paraconsistency, see my "Paraconsistent Logic."

As in the classical case, evaluations of propositional parameters are extended to all formulas by recursive conditions. The conditions for $\neg$ and $\wedge$ are as follows. (The conditions for $\vee$ are dual to those for $\wedge$, and may safely be left as an exercise.)

$R(\neg\alpha, 1)$ if and only if $R(\alpha, 0)$
$R(\neg\alpha, 0)$ if and only if $R(\alpha, 1)$
$R(\alpha \wedge \beta, 1)$ if and only if $R(\alpha, 1)$ and $R(\beta, 1)$
$R(\alpha \wedge \beta, 0)$ if and only if $R(\alpha, 0)$ or $R(\beta, 0)$

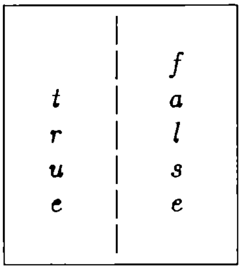
Thus, $\neg\alpha$ is true if and only if α is false, and vice versa. A conjunction is true if and only if both conjuncts are true; false if and only if at least one conjunct is false—all very familiar.

To complete the picture, we need a definition of logical consequence. This also presents no surprises. An inference is valid if and only if whenever the premises are true, so is the conclusion. Thus, if Σ is a set of formulas:

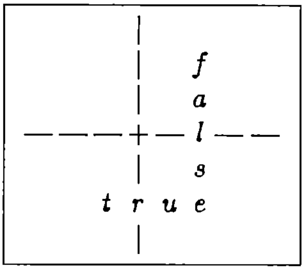
$\Sigma \models \alpha$ if and only if for all R (if $R(\beta, 1)$ for all $\beta \in \Sigma$, $R(\alpha, 1)$)

It is now easy to see why the logic is paraconsistent. Choose an evaluation, R, that relates p to both 1 and 0, but relates q only to 0. Then it is easy enough to see that both p and $\neg p$ (and $p \wedge \neg p$) are true under R (and false as well, but at least true), while q is not. Hence $p, \neg p \not\models q$. For future reference, note that the same evaluation refutes the disjunctive syllogism: $p, \neg p \vee q \vdash q$.

The logic given here should look very familiar. It *is* very familiar. It is exactly the same as classical logic, except that one does not make the assumption, usually packed into textbooks of logic without comment, that truth and falsity in an interpretation are exclusive and exhaustive. The difference between classical logic and the above logic can therefore be depicted very simply. In classical logic, each interpretation partitions the set of formulas as follows:



In the paraconsistent logic, an interpretation may partition in this way: classical interpretations are, after all, simply a special case. But in general, the partitioning looks like this:



The crucial question now is: Assuming that all the other assumptions packed into the story are right, should we, or should we not, countenance interpretations that correspond to the second picture? There is no quick way with this question. Each logic encapsulates a substantial metaphysical/semantical *theory*. It should be noted that a paraconsistent logician does not have to hold that truth itself behaves as in the second picture. They have to hold only that in defining validity one has to take into account interpretations that do. Although the claim that truth itself behaves like this is one argument for this conclusion, it is not the only one. If we think of interpretations as representing situations about which we reason, then interpretations of the second kind might be thought to represent "impossible" situations that are inconsistent or incomplete, such as hypothetical, counterfactual, or fictional situations, or as situations about which we have incomplete or inconsistent information. One may well suppose that there are, in some relevant sense, such situations, and that they play an important metaphysical and/or semantical role.

More boldly, one may suppose that truth itself behaves according to the second picture, and hence that there must be at least one interpretation that does, namely, the interpretation that assigns truth values in accord with the actual. One cannot simply *assume* that it does not. Here, again, lie profound metaphysical issues. Even the founder of logic, Aristotle, did not think that truth satisfies the first picture. According to him, statements about future contingents, such as the claim that there will be a sea battle tomorrow, are neither true nor false (unless you live in Bolivia).[4] The top left square of the second picture is therefore occupied. Modern logic has provided many other possible candidates for this square: statements employing non-denoting terms, statements about undecidable sentences in science or mathematics, category mistakes and other "nonsense," and so on.

---

[4] *De Interpretatione*, chapter 9. He seems to think that this is consistent with the law of excluded middle, however. At least, he defends this law in *Metaphysics* Γ.

The thought that the bottom right corner might also have denizens is one much less familiar to modern philosophers. Yet there are plausible candidates. Let me give two briefly.[5] The first concerns paradoxes of self-reference. Let us take the liar as an example. The natural and most obvious principle concerning truth is encapsulated in the T-schema: for any sentence, α: $T\langle\alpha\rangle \leftrightarrow \alpha$. I use '$T$' here as a truth predicate, and angle brackets as a name-forming device. With standard self-referential techniques, we can now produce a sentence, β, that says of itself that it is not true: $\neg T\langle\beta\rangle \leftrightarrow \beta$. Substituting β in the T-scheme and juggling a little gives $\beta \wedge \neg\beta$. Prima facie, then, β is a sentence that is both true and false, and so occupies the bottom right corner.

Another example: I walk out of the room; for an instant, I am symmetrically poised, one foot in, one foot out, my center of gravity lying on the vertical plane containing the center of gravity of the door. Am I in or not in the room? By symmetry, I am neither in rather than not in, nor not in rather than in. The pure light of reason therefore countenances only two answers to the question: I am both in and not in, or neither in nor not in. Thus, we certainly appear to have a denizen of either the top left or the bottom right quarter. But wait a minute. If I am neither in nor not in, then I am not (in) and not (not in). By the law of double negation, I am both in and not in. (Even without it, I am both not in and not not in, which is still a contradiction.) Hence we have a denizen of the bottom right.

There is, of course, much more to be said about both of these examples. But I do not intend to say anything further here.[6] The point is simply to illustrate some of the semantic/metaphysical issues that

[5] These and others are discussed at much greater length in my *In Contradiction* (Dordrecht: Nijhoff, 1987).

[6] Since the second example is not as familiar as the first, let me add one comment. Let us represent the sentence 'GP is in the room' by α. An obvious move at this point is to suggest that α is, in fact, a denizen of the top left quarter, but that one cannot express this fact by saying that I am neither in nor not in the room. What one has to say is that neither α nor its negation is true, $\neg T\langle\alpha\rangle \wedge \neg T\langle\neg\alpha\rangle$. This is certainly not an explicit contradiction. Unfortunately, it, too, soon gives one. The T-schema for α and $\neg\alpha$ tell us that $T\langle\alpha\rangle \leftrightarrow \alpha$ and $T\langle\neg\alpha\rangle \leftrightarrow \neg\alpha$. Contraposing and chaining together gives: $\neg T\langle\neg\alpha\rangle \leftrightarrow T\langle\alpha\rangle$, and we are back with a contradiction. A natural move here is to deny the T-schema for α or $\neg\alpha$ (presumably these stand or fall together). But on what ground can one reasonably do this? 'GP is in the room' is a perfectly ordinary sentence of English. It is meaningful, and so must have truth conditions. (In fact, most of the time it is simply true or false.) These (or something equivalent to them) are exactly what the T-schema gives. Compare this with the case of the liar. Many have been tempted to reject the T-schema for the liar sentence on the ground that the sentence is semantically defective in some way. No such move seems to be even a prima facie possibility in the present case.

must be hammered out even to decide whether truth itself satisfies the first or the second picture. To suppose that the answer is obvious, or that the issue can be settled by definition, is simple dogmatism.

There is a famous defense of classical logic by W. V. Quine which comes very close to this, in fact. Someone who takes there to be interpretations corresponding to the second picture just "does not know what they are talking about": to change the logic is to "change the subject." It is changing the subject only if one assumes in the first place that validity is to be defined in terms only of interpretations that satisfy the first picture—which is exactly what is at issue here. Two logicians who subscribe to different accounts of validity are arguing about the same subject, just as much as two physicists who subscribe to different accounts of motion.[7]

Now, finally, to return to the main point. I have not shown that explosion fails, that one ought to take into the scope of logic situations that are inconsistent and/or incomplete, though I do take it that when the dust settles, this will be seen to be the case, and that even truth itself requires the second picture.[8] The point of the above discussion is simply to show that the failure of explosion is a plausible logico-metaphysical view, and that one cannot simply *assume* otherwise without begging the question.

II OBJECTION TWO: CONTRADICTIONS CANNOT BE TRUE

Let us turn now to objection number two. This is to the effect that contradictions cannot be true. Since one ought to believe only what is true, contradictions ought not to be believed.

This argument appeals to the *law of noncontradiction* (LNC): nothing is both true and false. The first thing we need to do is distinguish clearly between the LNC and explosion, as they are very different. For a start, as we have seen, explosion is a relative newcomer on the logical scene; the LNC is not. It is true that some have challenged it: some Presocratics, such as Heraclitus; some Neoplatonists, such as Cusanus; and some dialecticians, such as G. W. F. Hegel. But since the time of Aristotle, it is a principle that has been very firmly entrenched in Western philosophy. (Its place in Eastern philosophy is much less secure.) The view that the LNC fails, that some contradictions are true, is called *dialetheism*. As we have already seen, one does not have to be a dialetheist to subscribe to the correctness of a paraconsistent logic, though if one is, one will. As we also saw, though, there are arguments that push us toward accepting dialetheism. Is there any reason why one should reject these a priori? Why, in other words, should we accept the LNC?

---

[7] For references to Quine, with further discussion, see my "On Alternative Geometries, Arithmetics and Logics: A Tribute to Łukasiewicz," in M Baghramian, ed., *Proceedings of the Conference Łukasiewicz in Dublin* (forthcoming).

[8] As a matter of fact, though, I think that its top left quarter is empty. See my *In Contradiction*, chapter 4.

The locus classicus of its defense is Aristotle's *Metaphysics* Γ4. It is a striking fact about the law that there has not been a sustained defense of it *since* Aristotle (at least, of which I am aware). Were his arguments so good that they settled the matter? Hardly. There are about seven or eight arguments in the chapter (it depends how you count). The first occupies half the chapter. It is long, convoluted, and tortured. It is not at all clear how it is supposed to work, let alone that it works. The other arguments in the chapter are short, often little more than throw-away remarks, and are at best, dubious. Indeed, most of them are clearly aimed at attacking the view that *all* contradictions are true (or even that someone can *believe* that all contradictions are true). Aristotle, in fact, slides back and forth between 'all' and 'some' with gay abandon. His defense of the LNC is therefore of little help.[9]

So what other arguments are there for the LNC? Very few of which I am aware, and none that survives much thought. Let me mention four here. The first two, some have claimed, are to be found in Aristotle. I doubt it, but let us not go into this here.

According to the first argument, contradictions have no content, no meaning. If so, then, a fortiori, they have no true content: contradictions cannot be true. The first thing to note about this objection is that it is not only an objection against dialetheism, but also against classical logic. For in classical logic, contradictions have *total* content, they entail everything. One who subscribes to orthodox logic cannot, therefore, wield this objection.

There have been some who endorsed different propositional logics, according to which contradictions do entail nothing, and so have no content.[10] But the claim that contradictions have no content does not stand up to independent inspection. If contradictions had no content, there would be nothing to disagree with when someone uttered one, which there (usually) is. Contradictions do, after all, have meaning. If they did not, we could not even understand someone who asserted a contradiction, and so evaluate what they say as false (or maybe true). We might not understand what could have brought a person to assert such a thing, but that is a different matter—and the same is equally true of someone who, in broad daylight, asserts the clearly meaningful 'It is night'.

---

[9] For a detailed analysis of Aristotle's arguments, see my "To Be *and* Not to Be—That Is the Answer: On Aristotle on the Law of Non Contradiction," *Philosophiegeschichte und Logische Analyse*, I (1998): 91-130.

[10] See my "Negation as Cancellation, and Connexivism."

A second objection (to be found, for example, in J. M. E. McTaggart) is to the effect that, if contradictions could be true, ***nothing*** could be meaningful. The argument here appeals to the thought that something is meaningful only if it excludes something (***omnis determino est negatio***): a claim that rules out nothing, says nothing. Moreover, it continues, if α does not rule out $\neg\alpha$, it rules out nothing. An obvious failing with this argument is, again, the slide from 'some' to 'all'. Violation of the LNC requires only that some statements do not rule out their negations (whatever that is supposed to mean). The argument depends on the claim that ***nothing*** rules out its own negation.

But there is a much more fundamental flaw in the argument than this. The premise that a proposition is not meaningful unless it rules something out is just plain false. Merely consider the claim 'Everything is true'. This rules nothing out: it entails everything. Yet it is quite meaningful (it is, after all, false). If you are in any doubt over this, merely consider its negation 'Something is not true'. This is clearly true—and so meaningful. How could a meaningful sentence have a meaningless negation?

A third argument for the LNC, and one that is typical of many, starts from the claim that the correct truth conditions for negation are as follows:

$\neg\alpha$ is true if and only if α is not true

Now, suppose that $\alpha \wedge \neg\alpha$ is true; then, assuming that conjunction behaves normally, α is true and $\neg\alpha$ is true. Hence by the truth conditions of negation, α is both true and not true, which is impossible.

It is not difficult to see what is wrong with this argument. For a start, the truth conditions of negation are contentious. (Compare them with those given in the previous section.) More importantly, why should one suppose that it is impossible for α to be both true and not true? Because it is a contradiction. But it is precisely the impossibility of having true contradictions for which we were supposed to be arguing. The argument, therefore, begs the question, as do many of the other arguments of which I am aware.[11]

The fourth, and final, argument I shall mention is an inductive one. As we review the kinds of situations that we witness, very few of them would seem to be contradictory. Socrates is never both seated and not seated; Brisbane is firmly in Australia and not not in it. Hence, by induction, no contradictions are true. Note that one does not have to suppose that logical principles are a posteriori for this form of argument to work. One can collect a posteriori evidence even for a priori principles. For example, one verifies $\alpha \vee \neg\alpha$ every time one verifies α.

---

[11] In particular, one may argue for the LNC from explosion, assuming that not all contradictions are true. But an appeal to explosion would beg the question, as we have already seen.

The flaws of this argument are apparent enough, though. It is all too clear that the argument may be based on what Ludwig Wittgenstein called "an inadequate diet of examples." Maybe Socrates *is* both sitting and not sitting sometimes: at the instant he rises. This, being instantaneous, is not something we observe. We can tell it to be so only by a priori analysis. Worse, counterexamples to the principle are staring us in the face. Think, for example, of the liar. Most would set an example such as this aside, and suppose there to be something wrong with it. But this may be short-sighted. Consider the Euclidean principle that the whole must be larger than its parts. This principle seemed to be obvious to many people for a long time. Apparent counterexamples were known from late antiquity: for example, the set of even numbers appeared to be the same size as the set of all numbers. But these examples were set aside, and just taken to show the incoherence of the notion of infinity. With the nineteenth century all this changed. There is nothing incoherent about this behavior at all: it is paradigmatic of infinite collections. The Euclidean principle holds only for finite collections; and people's acceptance of it was due to a poor induction from unrepresentative cases. In the same way, once one gets rid of the idea, in the form of explosion, that inconsistency is incoherent, the liar and similar examples can been seen as paradigm citizens of a realm to which our eyes are newly opened (we can call it, by analogy with set theory, the *transconsistent*). In any case, the inductive argument to the LNC is simply a poor one.

It is sometimes said that dialetheism is a position based on sand. In fact, I think, it is quite the opposite: it is the LNC that is based on sand. It appears to have no rational basis; and the historical adherence to it is simply dogma. Hence—and finally to return to the second objection—it fails.

III. OBJECTION THREE: CONTRADICTIONS CANNOT BE BELIEVED RATIONALLY

The third objection is that, even if contradictions could be true, they cannot be believed rationally, consistency being a constraint on rationality; hence one ought not to believe a contradiction since this would be irrational.

We have already seen, in answer to the first objection, that this objection fails. The paradox of the preface shows that it can be quite rational to have inconsistent beliefs. Hence, consistency is not an absolute constraint on rationality. The rational person apportions their beliefs according to the evidence; and if the evidence is for inconsistent propositions, so be it.

There is, of course, more to the story than this. To approach it, let me take what will appear to be a digression for a moment. Have you ever talked to a flat-earther, or someone with really bizarre religious beliefs—not one who subscribes to such a view in a thoughtless way, but someone who has considered the issue very carefully? If you have, then you will know that it is virtually impossible to show their view to be wrong by finding a knock-down objection. If one points out to the flat-earther that we have sailed round the earth, they will say that one has, in fact, only traversed a circle on a flat surface. If one points out that we have been into space and seen the earth to be round, they will reply that it only appears round, and that light, up there, does not move in straight lines, or that the whole space-flight story is a Central Intelligence Agency put-up, and so on. In a word, their views are perfectly consistent. This does not stop them from being irrational, however. How to diagnose their irrationality is a nice point, but I think that one may put it down to a constant invoking of ad hoc hypotheses. Whenever one thinks one has a flat-earther in a corner, new claims are pulled in, apparently from nowhere, just to get them out of trouble.

What this illustrates is that there are criteria for rationality other than consistency, and that some of these are even more powerful than consistency. The point is, in fact, a familiar one from the philosophy of science. There are many features of belief that are rational virtues, such as simplicity, problem-solving ability, nonadhocness, fruitfulness, and, let us grant, consistency. These criteria are all independent, however, and may even be orthogonal, pulling in opposite directions. Now, what should one do if, for a certain belief, all of the criteria pull toward acceptance, except consistency—which pulls the other way? It may be silly to be a democrat about this and simply count the number of criteria on each side; but it seems natural to suppose that the combined force of the other criteria may trump inconsistency. In such a case, then, it is rational to have an inconsistent belief.

The situation I have outlined is an abstract one; but it seems to me that it, or something like it, already obtains with respect to theories of truth. Since the abstract point is already sufficient answer to the objection with which we are dealing, I do not want to defend the example in detail here; still, it will serve to put some flesh on the abstract bones. The following is a simple account of truth. Truth is a principle that is characterized formally by the *T*-schema: for every sentence, α, $T\langle\alpha\rangle \leftrightarrow \alpha$ (for a suitable conditional connective)—and "that's an end on't." (There may be more to be said about truth, but nothing that can be captured in a formalism.) This account is inconsistent: when suitable self-referential machinery is present (say, in the form of arithmetic), the liar paradox is forthcoming. Yet the inconsistencies are isolated. In particular, it can be shown that, when things are suitably set up, inconsistencies do not percolate into the purely arithmetic machinery. In fact, it can be shown that any sentence that is grounded (in Saul Kripke's sense) behaves consistently.[12] What are the alternatives to such an account? There is a welter of them: Alfred Tarski's, Kripke's, Anil Gupta and Hans Herzberger's, Jon Barwise and John Perry's, Vann McGee's, et alia. These may all have the virtue of consistency, but the other virtues are thinly distributed among them. They often have strong ad hoc elements; they are complex, usually involving transfinite hierarchies; they have a tendency to pose just as many problems as they solve; and it is not clear that, in the last instance, they really solve the problem they are supposed to: they all seem subject to extended paradoxes of some kind.[13] It seems to me that rationality speaks very strongly in favor of the simple inconsistent theory. This is exactly a concrete case of the abstract kind I have described.

Naturally, it may happen that someone, a hundred years hence, will come up with a consistent account of truth with none of these problems, in which case, what it is rational to believe may well change. But that is neither here nor there. Rational belief about anything is a fallible matter. It is a mistake to believe where the evidence does not point; but it is equally a mistake not to believe where the evidence points.

I have argued that it may well be rational to believe a contradiction, and shown how this may arise. If there is sufficient evidence that something is true, one ought, rationally, to accept it. Let me consider just one reply. It is natural to suppose that there is a dual principle here: if there is sufficient evidence that something is false, one ought, rationally, to reject it. If, therefore, there is strong evidence that contradictories, α and $\neg\alpha$, are both true, there is evidence that both are also false. One ought, then, to reject both.

No. In the appropriate sense, truth trumps falsity. Truth is, by its nature, the aim of cognitive processes, such as belief. (This is the 'more' to truth to which I referred above.) It is constitutive of truth that that is what one ought to accept. Falsity, by contrast, is merely truth of negation. It has no independent epistemological force. One should not necessarily, therefore, reject something simply because its negation turns out to be true.

---

[12] For a proof of this, see my "Paraconsistent Logic," section 8.

[13] See my *In Contradiction*, chapter 1.

The situation may well be different with respect to untruth. At least arguably, if something is shown not to be true then one ought to reject it.[14] But one cannot suppose that falsity and untruth are the same thing, if the second picture drawn in connection with objection number one is correct. If one does so suppose, as epistemologists traditionally have done, then something shown to be false is shown to be untrue, and so not a target for belief. This may be why the dual principle has its appeal. But once one sees that truth and falsity (that is, truth of negation) cannot always be separated, like the elements of a constant-boiling mixture, it becomes clear that this is overly simplistic. At any rate, we have seen more than enough to answer objection number three.

IV OBJECTION FOUR: IF CONTRADICTIONS WERE ACCEPTABLE, PEOPLE COULD NEVER BE RATIONALLY CRITICIZED

The fourth objection also concerns rationality, and is to the effect that, if contradictions were acceptable, no one could be rationally criticized for the views that they hold. The thought here is that, if you hold some view, and I object to it, there is nothing, rationally, to stop you from maintaining both your original view *and* my objection.

The most obvious failing of this argument is that it makes the familiar and illicit slide from 'some' to 'all'. The mere fact that some contradictions are rationally acceptable does not entail that all are. The charge 'You accept some contradictions to be true, so why should you not believe any contradiction to be so?' is as silly as the charge 'You believe something to be true, so why should you not believe anything to be so?'

It might be argued that, if it is logically possible for any contradiction to be true (as it is in the semantics we looked at in reply to objection one), then all contradictions are rationally acceptable. This, though, most certainly does not follow either. The fact that something is a logical possibility does not entail that it is rational to believe it. It is logically possible that I am a fried egg, though believing that I am is ground for certifiable insanity. As we saw in reply to the last objection, there is a lot more to rationality than consistency. A view, such as that the earth is flat, may be quite consistent (and so logically possible in traditional terms), and yet quite irrational.

A person's views may be rationally criticized if they can be shown to entail something that is rationally unacceptable. This might be a contradiction, but it might be some noncontradiction. Some noncontradictions, for example, that I am a fried egg, are, in fact, better than some contradictions, for example, that the liar sentence is both true and false. In the last instance, what is rationally acceptable, and what is not, is likely to be a holistic matter, to be determined by the sort of criteria I discussed in response to the last objection. Let me illustrate again. I argued there that an inconsistent account of truth, which endorsed the *T*-schema, was preferable to the numerous consistent accounts available. Suppose that it turned out, in defending the inconsistent view, that it had to be shorn up in the same methodologically unsatisfactory ways as extant consistent accounts—for example, to avoid strengthened paradoxes—until it was just as complex and contrived. It would then cease to be rational to accept it. The fact that one can accept some contradictions would do nothing to help the matter.

---

[14] One may well contest this, too, see my "Can Contradictions Be True? II," *Proceedings of the Aristotelian Society*, Supplementary Volume LXVII (1993): 35-54.

This is a perfectly adequate reply to the objection, but let me say a little more. I am frequently asked for a criterion as to when contradictions are acceptable and when they are not. It would be nice if there were a substantial answer to this question—or even if one could give a partial answer, in the form of some algorithm to demonstrate that an area of discourse is contradiction free. But I doubt that this is possible. Nor is this a matter for surprise. Few would now seriously suppose that one can give an algorithm—or any other informative criterion—to determine when it is rational to accept something. There is no reason why the fact that something has a certain syntactic form—be it $p \wedge \neg p$ or anything else—should change this. One can determine the acceptability of any given contradiction, as of anything else, only on its individual merits.

Despite this, I do think that there are general reasons as to why contradictions are a priori improbable. Classical logicians, who hold that contradictions all have probability 0, should agree with this! But it may reasonably be asked why one should suppose this to be so, once one has given up the assumption that that probability is 0. The answer to this question is simply that the statistical frequency of true contradictions in practice is low. This low frequency suffices to determine a low probability.

How do we know that true contradictions have a low frequency? Return to the inductive argument for the LNC which we considered in connection with objection number two. I pointed out there how weak this was as an argument for the *universality* of contradiction freedom. But as an argument for the infrequency of contradictions it is much better. The counterexamples to the universality of the LNC are of very particular sorts (involving self-reference, or states of affairs that are but instantaneous, and so on), and we do not deal with these kinds of situations very often.

As a measure of this fact, recall the disjunctive syllogism (α, $\neg\alpha \vee \beta \vdash \beta$). This is not valid in the semantics we looked at. Yet we use it all the time in practice, and rarely does it lead us astray. It will lead us astray only when there is a situation where α is both true and false, and β is not true. Hence, there are few such situations. This could be for two reasons. The first is that there are few αs that are both true and false; the second is that there are few βs that are not true. But we may rule out the second possibility: if this were the case, then we would rarely go wrong in *any* conclusion we draw, but we do. Hence the frequency of true contradictions is low.

The fact that contradictions have low probability grounds the fact that inconsistency is a rational black mark. If we have views that are inconsistent then we are probably incorrect. We should go back and examine why we hold such a view, and what the alternatives are. We may find that we would be better off going a different way. But we may find that there are no better ways to go. In which case, we may just have to conclude that the improbable is the case. After all, the improbable happens sometimes. We would seem to be in exactly this situation with respect to theories of truth and the liar. In one way or another, we have been over this ground for over 2,000 years—for the last one hundred years very intensively—and no satisfactory consistent theory has been found. At any rate, inconsistency provides a prima facie ground for rejecting a view. One cannot simply accept a contradiction. There is other work to be done. This provides another answer to objection number four.[15]

V OBJECTION FIVE: IF CONTRADICTIONS WERE ACCEPTABLE, NO ONE COULD DENY ANYTHING

The final objection takes us into new territory, one concerning public speech. The argument here is to the effect that, if contradictions were acceptable, then no one would have a way of denying anything: whenever they asserted $\neg\alpha$, this would not show that they rejected α, for they might accept both α and $\neg\alpha$.

To discuss this argument, we first need to be clear about asserting and denying.[16] These are speech acts, like questioning or commanding. Which ones? If I assert something, α, then this is a speech act whose intention is to get the hearer to believe α, or at least, believe that I believe α—with whatever Gricean sophistication one may wish to add. If I deny something, α, then this is a speech act whose intention is to get the hearer to reject α (cast it out from their beliefs, and/or refuse to accept it), or at least, to get the listener to believe that I reject it—with whatever Gricean sophistication one may wish to add.

---

[15] For a further discussion of the issue, see my *In Contradiction*, chapters 7 and 8.

[16] The following follows my "Can Contradictions Be True? II." The discussion is taken further in my "What Not? A Dialetheic Account of Negation," in Gabbay and H. Wansing, eds , *Negation* (Dordrecht: Kluwer, forthcoming).

Now, prima facie, at least, assertion and denial are quite distinct kinds of speech act, and this is the way they have often been understood traditionally—for example, in the Port Royale logic (though, of course, the point was not put in terms of speech acts, which is a modern invention). But Frege suggested, and many now accept, that denial may be reduced to assertion by the equation:

denial = assertion of negation

This identity is incorrect. To assert the negation of something is not necessarily to deny it. When I, for example, assert the negation of the liar sentence, I am not denying it. After all, I accept it, and intend you to do the same. Nor does this really have anything to do with dialetheism. We, all of us, discover sometimes—maybe by the prompting of some Socratic questioner—that our beliefs are inconsistent. We assert α, and then a little later assert $\neg\alpha$. We may well wish to revise our views in the light of this—we usually do. But that is not the point here. The point is simply that in asserting $\neg\alpha$, we are not denying α. We do accept α; that, after all, is the problem. Hence, to assert a negation is not necessarily to deny—and the problem that this objection points to is just as much a problem for the classical logician as for the dialetheist.

More importantly, and conversely, one can deny something without asserting a negation. One can use a certain tone of voice or body language (like thumping the table). The issue is simply one of how to convey one's intentions. This is the solution to the problem. In fact, one can often deny something by asserting its negation. (Thus, this objection, again, makes the now very familiar slide from 'some' to 'all'.) Whether or not one is denying just depends. This raises the question of how one knows whether someone who utters a negated sentence is asserting or is denying. I doubt that there is any simple way of answering this question. In any case, it is of a kind very familiar from speech-act theory. Someone utters 'The door is open'. This could be an assertion, a question, a command. How does one know? Well, one has to determine the utterer's intentions; to do this one needs to know all kinds of things about language, the context, the social power relations, and the like. Never mind if we do not know exactly how we do it. We do it all the time.

Before we leave the subject, let me mention one final, related point. It is sometimes said that it is impossible even to express contradictory beliefs: if someone asserts α, and then asserts $\neg\alpha$, they have not expressed contradictory beliefs; their second utterance merely "cancels out" the first.[17] This could be an appeal to the claim that contradictions have no content, with which I have already dealt. But more likely, it is an appeal to the idea that asserting a negation is a denial. To deny something asserted is to "cancel out" the assertion, in the sense that it leaves the hearer no coherent way of interpreting the utterer's beliefs, short of supposing that they have changed their mind. But as we have seen, uttering a negation may just be a simple assertion: there need be no cancellation of any kind.

The ambiguity of 'assertion' (between the content of what is asserted, and the act of assertion) bedevils the history of logic until Frege. The ambiguity of 'denial' (between the content of a negated sentence, and an act of denial) may still bedevil it, as objections of the kind we have been looking at demonstrate.[18]

VI. CONCLUSION

We have now considered all the supposed a priori objections I started by enumerating. The sophist Gorgias argued that there is no truth; and even if there were, you could not know it; and even if you could, you could not express it. The arguments we have been looking at might be summarized, loosely, by saying, similarly: a contradiction cannot be true; but even if it could be, you could not know it; and even if you could, you could not express it. The arguments, as we have seen, have no more force than Gorgias's arguments. So what is so bad about contradictions? Maybe nothing.

GRAHAM PRIEST

University of Queensland

---

[17] For example, Peter Strawson runs this line. For references and further discussion, see my "To Be *and* Not to Be—That Is the Answer: On Aristotle on the Law of Non Contradiction," section 13.

[18] I have heard it suggested that once one distinguishes between negation and denial, there will be versions of the liar paradox, formulated in terms of denial, that a paraconsistent solution cannot handle. This is not so. The standard liar is a sentence, α, of the form $\neg T\langle\alpha\rangle$. Let us write $\dashv$ as a force operator, indicating denial. The analogue would be a sentence, α, such that α is $\dashv T\langle\alpha\rangle$. But this makes no sense, since $\dashv$ is not part of a propositional content. We can formulate a proposition, α, whose content is 'I deny that α'. Does this pose problems? Well, if I deny it, then it is true, and presumably obviously so to me. So I ought not to deny it. Conversely, if I do not deny it, then it is false, and again, presumably, obviously so to me. So I ought to deny it. In either case, then, I am going to fail an obligation. Perhaps, in the end, one just has to live with this fact. It is not a contradiction (and even if it were, isolated contradictions need not be a problem for a dialetheist). Moreover, the dialetheist does not even have to agree with the argument. As we have already seen in reply to objection number three, it is not necessary to reject (and so deny) something simply because it is false. A classical logician, on the other hand, for whom this is just as much a problem, cannot make the same move. Note that there are other paradoxes in the vicinity here that are even more embarrassing for a classical logician; see my "Gaps and Gluts: Reply to Parsons," *Canadian Journal of Philosophy*, XXV (March 1995): 57-66, especially section 4.

BOOK REVIEWS

Rationality, Allocation, and Reproduction. VIVIAN WALSH. New York: Oxford University Press, 1996. 304 p. Cloth $65.00.

Vivian Walsh brings together in this book the moral, methodological, historical, and economic concerns that have occupied him during his long and distinguished career. The central issue is the mismatch between everyday notions of rationality, which are informal and evaluative, and the notion of rationality which is central to neoclassical economics, which is formal and purportedly value free. But the book ranges much more widely. The first five chapters are concerned with rationality. Chapters six and seven discuss general competitive equilibrium and its Pareto optimality. Chapters eight and nine explore the notion of equilibrium in recent neo-Ricardian models and the difficulties of incorporating demand and dynamics into them.

This is an erudite book with a large bibliography, and it sketches many positions found in the literature on economics and philosophy. The erudition sometimes impedes the argument. Readers looking for a unified defense of a precisely delineated thesis will be disappointed, but readers who are willing to examine individually the many strands of Walsh's discussion will find much that is insightful.

I shall criticize three related theses Walsh defends. The first and most central is that the standard theory of economic rationality—utility theory—commits economists to a weak form of utilitarianism. Consequently (second), economics cannot be a "value-free" science. Third, the obsession of neoclassical economists with the notion of Pareto optimality derives from this weak utilitarianism and excludes all nonwelfarist moral considerations. Although I have defended apparently similar views myself,[1] I shall argue that these theses involve serious mistakes.

Neoclassical economics (the standard economics of today's textbooks, which first appeared in the 1870s in the works of William Stanley Jevons, Carl Menger, and Léon Walras) owes a debt to utilitarianism. Jevons conceived of utility much as Jeremy Bentham did, as a mental state (happiness) or a property of objects causing this

---

[1] See Michael S. McPherson and my *Economic Analysis and Moral Philosophy* (New York: Cambridge, 1996).

0022-362X/98/9508/427–30

mental state. Individuals prefer *X* to *Y* *because* *X* yields more utility. Francis Ysidro Edgeworth, Vilfredo Pareto, and then more decisively John Hicks and R. G. D. Allen in the 1930s showed how to free the theory of choice from any commitment to hedonistic psychology. Their theory of choice requires instead that agents have complete and transitive preference rankings that determine how they choose. (The representation of such rankings by continuous utility functions requires, however, a further continuity condition.) Walsh sees Hicks and Allen as spearheading "a systematic effort...to banish what was regarded as the substance of utilitarianism from neo-classical economics" (35).

Although Walsh's view has some historical backing, he does not distinguish between (a) a hedonist theory of choice, (b) a hedonist theory of welfare, and (c) a moral theory, such as utilitarianism, that evaluates outcomes, institutions, and actions in terms of human welfare. Hicks and Allen show only how to formulate economics without relying on a hedonist theory of choice, yet Walsh maintains that their work led "to the unwarranted belief that neo-classical theory had been purged of all elements of utilitarianism" (36).

Walsh maintains that this project failed because even though ordinal-utility theory rules out any summation of utilities, it "is nevertheless recognizably utilitarian by virtue of being teleological (or consequentialist) and welfarist" (39). This conclusion is mistaken. Utility theory might appear "consequentialist" because it holds that individuals maximize utility, but to maximize utility is merely to do what one most prefers to do. Utility theory by itself places no constraints on the objects of preference. In offering an alternative to a hedonist theory of choice, it says nothing about welfare and nothing about moral evaluation. Modeling individual choice as utility maximization or even going on (independently) to identify welfare with the satisfaction of preference does not commit one to evaluating policies in terms of welfare.

The second thesis, that neoclassical economics cannot be "value free," is less radical than it might appear, because Walsh apparently equates neoclassical economics with neoclassical welfare theory. He never discusses whether, for example, a model used to estimate revenue from a proposed tax could be value free. His claim that neoclassical welfare economics is not value free seems to derive from his view of the moral content of the theory of rationality. He accuses economists of believing inconsistently that claims about rationality have no moral implications even though they imply the existence of an equilibrium that is "superior" to alternative outcomes (131). Indeed, though naming no names, Walsh maintains that economists "believed that claims concerning the superiority of 'Pareto optimal' over 'non-Pareto optimal' equilibria were not ethical claims" (185). No doubt some economists (like some philosophers and some plumbers) have said silly things, but one would like some details. One possibility is the one Walsh defends: utility theory is a moral theory that implies moral evaluations of equilibria, but economists fail to recognize the moral content. A second possibility is that utility theory has no moral content and that neither do these specific claims about the superiority of Pareto optima. If *A* is a Pareto optimum, then there is no alternative that satisfies somebody's preferences better and nobody's preferences worse. Rather than believe that theses concerning the moral superiority of *A* to *B* are not moral claims, economists may have believed that the assertion that Pareto optima are superior *at satisfying preferences* is not a moral claim. From the theory of rationality and other apparently nonevaluative premises, the first welfare theorem proves that there exists a Pareto-optimal competitive general equilibrium. To prove this theorem and to take it as showing (in the sense just specified) the superiority of competitive equilibrium apparently involves no leap from facts to values.

I may misunderstand what Walsh means, because he takes the distinction between facts and values to presuppose a commitment to logical positivism (137). Philosophers have, he maintains, abandoned the distinction (137). Most philosophers have indeed abandoned the view that ethical utterances are meaningless grunts of approval or disapproval, and most are aware that there is no simple and sharp distinction between normative and non-normative claims. Perhaps this is all that Walsh means. No one maintains that welfare economics is entirely value free, and Walsh is right to hold that economists slide easily from technical optimality claims to moral evaluations.

Finally, Walsh argues that economists are fixated on Pareto optimality because of the weak utilitarianism built into their theory of rationality. They are, he maintains, consequently unable to take into account nonwelfarist moral considerations (186). Part of what he maintains is correct. Even though utility theory does not *imply* a welfarist moral theory, many economists are nevertheless welfarists. In addition, most neoclassicals reject interpersonal utility comparisons and so cannot be classical utilitarians. Since many economists identify welfare with the satisfaction of preferences, they will conclude that in a non-Pareto-optimal state of affairs someone is needlessly badly off—it is possible to make someone better off without making anyone worse off. Finally, if economists accept an uncontroversial principle of minimal benevolence—that, other things being equal, states of affairs are better if people are better off, then they have established that Pareto optima have some moral credentials.

There is, however, nothing in this argument or in neoclassical welfare economics which commits one to adopting the Pareto criteria as the *sole* method of evaluating economic institutions and outcomes, and nobody defends such a view. When Walsh argues that the Pareto optimality of a state of affairs is not a sufficient basis for judging it favorably, he is tackling a straw man. The standard view is, in contrast, that Pareto optimality concerns *efficiency* (in satisfying preferences). Indeed, many economists think that the word 'optimality' is misleading, and that one should talk about Pareto efficiency instead. Textbooks in welfare economics maintain that the evaluation of outcomes and institutions depends both on considerations of efficiency and on other factors, such as equity. When it comes to doing applied welfare economics, these other moral concerns are unfortunately frequently forgotten, and discussions of efficiency often occupy not only center stage but the whole of the stage. But Walsh is not criticizing such lapses. He holds that standard welfare economics "enthrones one particularly narrow moral claim—the Pareto criterion—which is not even seen as a moral claim, at the expense of all non-welfarist or non-consequentialist moral claims..." (186). In fact, in regarding Pareto efficiency as just one moral consideration among others (in contrast to the utilitarian who admits no ultimate consideration apart from maximizing total utility), welfare economics leaves open a space for considerations of equity, equality, freedom, rights, virtues—you name it.

In practice, welfare is the name of the game, and Pareto improvements and potential Pareto improvements are pretty much all with which standard welfare economics concerns itself. Walsh is right to hold that normative economics leaves a great deal to be desired and to deny that utility theory provides an adequate theory of rationality. The details of his criticisms are, however, faulty.

DANIEL M. HAUSMAN

University of Wisconsin/Madison

THE JOURNAL OF PHILOSOPHY

VOLUME XCV, NO. 9, SEPTEMBER 1998

HAVING THE WORLD IN VIEW: SELLARS, KANT, AND INTENTIONALITY*

LECTURE I: SELLARS ON PERCEPTUAL EXPERIENCE

In his seminal set of lectures, "Empiricism and the Philosophy of Mind,"[1] Wilfrid Sellars offers (among much else) the outlines of a deeply Kantian way of thinking about intentionality—about how thought and language are directed toward the world. Sellars describes *Science and Metaphysics: Variations on Kantian Themes*,[2] his major work of the next decade, as a sequel to "Empiricism and the Philosophy of Mind" (*ibid.*, p. viii). The later work makes explicit the Kantian orientation of the earlier; Sellars now shows a conviction that his own thinking about intentionality (and, indeed, about everything) can be well expounded through a reading of Kant. I do not think it is far-fetched to attribute to Sellars a belief on the following lines: no one has come closer than Kant to showing us how to find intentionality unproblematic, and there is no better way for us to find intentionality unproblematic than by seeing what Kant was driving at. That means rethinking his thought for ourselves and, if necessary, correcting him at points where we think we see more clearly than he did what he should have been doing. Sellars does not hesitate to claim, on some points, to have a better understanding of the requirements of Kantian thinking than Kant himself achieved.

---

* A revised version of the Woodbridge Lectures, sponsored by the Department of Philosophy, Columbia University, on April 15, 16, and 17, 1997.

[1] In Herbert Feigl and Michael Scriven, eds., *Minnesota Studies in the Philosophy of Science*, Volume 1 (Minneapolis: Minnesota UP, 1956); reprinted (with some added footnotes) in Sellars's *Science, Perception and Reality* (London: Routledge, 1963; reissued, Atascadero, CA: Ridgeview, 1991) I shall give page references to the original printing, followed in parentheses by references to the reprinted version.

[2] London: Routledge, 1967; reissued, Atascadero, CA: Ridgeview, 1992.

0022-362X/98/9509/431–90

Now, I share this belief I have read into Sellars, that there is no better way for us to approach an understanding of intentionality than by working toward understanding Kant. I also believe that coming to terms with Sellars's sustained attempt to be a Kantian is a fine way to begin appreciating Kant, and thereby—given the first belief—to become philosophically comfortable with intentionality. I mean this as a partly backhanded compliment to Sellars. He makes perfectly clear how he thinks he has to correct Kant, and I want to suggest that the divergence is revealing. I think a fully Kantian vision of intentionality is inaccessible to Sellars, because of a deep structural feature of his philosophical outlook. I believe we can bring into clearer focus the way Kant actually thought about intentionality, and thereby—given that first belief—how we ourselves ought to think about intentionality, by reflecting on the difference between what Sellars knows Kant wrote and what Sellars thinks Kant should have written.[3]

The reading of Kant that I aim to sketch in these lectures is under construction in a collaborative enterprise I am privileged to be engaged in with my colleagues James Conant and John Haugeland. Here I want to make a standard prefatory remark, which I mean in a less ritualistic manner than is perhaps usual. Conant and Haugeland should receive full credit for anything in what follows that is helpful toward the understanding of Kant, and thereby toward the understanding of intentionality. The blame for anything unhelpful, or simply wrong, is mine alone. In particular, Conant and Haugeland should not be held responsible for the perhaps perverse idea that we can approach an understanding of Kant through seeing how close Sellars comes to Kant's picture; nor should they be held responsible for the details of my reading of Sellars.[4]

---

[3] It is a measure of how difficult it is to come to terms with Kant that this sort of indirect approach can be helpful. Dieter Henrich describes my references to Kant, in my earlier engagement with the issues I shall be considering in these lectures (*Mind and World* (Cambridge: Harvard, 1994; reissued with an Introduction, 1996)), as "platitüdinnahen"—"Zwei Naturalismen auf englisch," *Merkur*, DLXV (1996): 334-43 No doubt it is nearly platitudinous that sensibility must have a central role in any even approximately Kantian attempt at making intelligible the very idea of intentionality, the directedness of subjective states or episodes toward objects. But that is nearly platitudinous just because it is neutral between Sellars's reading of Kant and the quite different picture I was trying to give. Sellars thinks a properly Kantian position requires that conceptual episodes occur in perception in a way that is guided by "sheer receptivity." I do not believe that is a correct picture of the transcendental role of sensibility in a properly Kantian position. If this belief were platitudinous, Sellars could not have understood Kantian thinking as he does. I hope this will become clearer in these lectures.

[4] I have also benefited from years of fruitful exchange with Robert Brandom, and from his very helpful comments on a draft of these lectures.

I

Sellars's master thought in "Empiricism and the Philosophy of Mind" is this. There is a special category of characterizations of states or episodes that occur in people's lives—for instance, characterizations of states or episodes as *knowings*; and, we might add, corresponding characterizations of the people in whose lives the states or episodes occur—for instance, characterizations of people as *knowers*. In giving these characterizations, we place whatever they characterize in "the logical space of reasons" (*op. cit.*, §36, pp. 298-99 (169)). Sellars's thesis is that the conceptual apparatus we employ when we place things in the logical space of reasons is irreducible to any conceptual apparatus that does not serve to place things in the logical space of reasons. So the master thought as it were draws a line: above the line are placings in the logical space of reasons, and below it are characterizations that do not do that.

That is a merely negative specification of what we must distinguish from placings in the logical space of reasons. But Sellars is concerned to warn against a particular philosophical pitfall, namely, the temptation to suppose, of certain specific below-the-line characterizations, that they can fulfill tasks that can, in fact, be fulfilled only by above-the-line characterizations. This temptation is urgent in respect of some, in particular, of the characterizations that function below Sellars's line, and we need a positive specification of the characterizations that activate the temptation. Sellars sometimes suggests this helpful way of putting his thought: characterizations that affirm *epistemic* facts need to be distinguished from characterizations that affirm *natural* facts.[5] In these terms, his central thesis is that we must not suppose we can understand epistemic states or episodes in terms of the actualization of merely natural capacities—capacities that their subjects have at birth, or acquire in the course of merely animal maturation. I think 'epistemic' here amounts to something like 'concept involving'; I shall justify this interpretation shortly.

Assuming this interpretation for the moment, we can bring Sellars's thought into direct contact with Kant. The logical space of reasons, on this reading, is the logical space in which we place episodes or states when we describe them in terms of the actualization of conceptual capacities. What corresponds in Kant to this image of the

CALCUTTA UNIVERSITY
ALIPUR
CAMPUS

---

[5] See §17, p. 274 (146), and note the echo of §5, p. 257 (131), where Sellars warns of a mistake in epistemology that is "of a piece with the so-called 'naturalistic fallacy' in ethics." At §36, pp. 298-99 (169), he contrasts placing things in the space of reasons with "empirical description"; I think this formulation is less helpful.

logical space of reasons is the image of the realm of freedom. The way to understand the correspondence is to focus on the Kantian idea that conceptual capacities are essentially exercisable in judging. It is true, and important, that judging is not the only mode of actualization of conceptual capacities; I shall be exploiting the point in these lectures. But even so, judging can be singled out as the paradigmatic mode of actualization of conceptual capacities, the one in terms of which we should understand the very idea of conceptual capacities in the relevant sense. And judging, making up our minds what to think, is something for which we are, in principle, responsible—something we freely do, as opposed to something that merely happens in our lives. Of course, a belief is not always, or even typically, a result of our exercising this freedom to decide what to think. But even when a belief is not freely adopted, it is an actualization of capacities of a kind, the conceptual, whose paradigmatic mode of actualization is in the exercise of freedom that judging is. This freedom, exemplified in responsible acts of judging, is essentially a matter of being answerable to criticism in the light of rationally relevant considerations. So the realm of freedom, at least the realm of the freedom of judging, can be identified with the space of reasons.

Sellars describes the logical space of reasons as the space "of justifying and being able to justify what one says."[6] We can see this as a distinctively twentieth-century elaboration of a Kantian conception: the conception of a capacity to exercise, paradigmatically in judgment, a freedom that is essentially a matter of responsiveness to reasons. The twentieth-century element is the idea that this capacity comes with being initiated into language.

II

At a pivotal point in "Empiricism and the Philosophy of Mind," Sellars addresses the question whether empirical knowledge has foundations (*op. cit.*, §VIII, pp. 293-300 (164-70)). His answer is nuanced.

In an empiricistic foundationalism of the usual kind, it is not just that the credentials of all knowledge are ultimately grounded in knowledge acquired in perception. Beyond that, the grounding perceptual knowledge is atomistically conceived. Traditional empiricists take it that each element of the grounding knowledge can in principle be acquired on its own, independently not only of other elements of the grounding perceptual knowledge but also of anything

[6] §36, p. 299 (169). This connects with the perhaps infelicitously labeled thesis of "psychological nominalism": see §29, p. 289 (160); §31, p. 291 (162); and, for an anticipation early in the lectures, §6, p. 258 (131).

in the world view that is grounded on this basic stratum of knowledge.

What Sellars objects to in traditional empiricism is just this supposed independence. He writes:

> There is clearly *some* point to the picture of human knowledge as resting on a level of propositions—observation reports—which do not rest on other propositions in the same way as other propositions rest on them. On the other hand, I do wish to insist that the metaphor of "foundation" is misleading in that it keeps us from seeing that if there is a logical dimension in which other empirical propositions rest on observation reports, there is another logical dimension in which the latter rest on the former (*op. cit.*, §38, p. 300 (170); cf. §19, p. 275 (148)).

Sellars does not deny that there is a logical dimension in which observation reports are basic. His point is just to insist on the other logical dimension, in which observation reports depend on the world view that is grounded on them, in the logical dimension to which a traditional empiricism restricts itself. The result is a picture that is still in a way empiricist, by virtue of its acknowledgment of the one logical dimension, though it is separated from traditional empiricism by virtue of its insistence on the other.

Of course Sellars's point here is at least partly epistemological, in an intelligibly narrow sense; he is telling us how we should conceive the credentials in virtue of which a world view counts as knowledgeably held. But the divergence from traditional empiricism means that we cannot take Sellars to be doing epistemology in some sense that contrasts with reflecting about intentionality. It is indeed perceptual *knowledge* (knowledge expressed in observation reports) about which he is here urging that it depends, in respect of the concepts that figure in it, on a world view. But that is just a case of something more general; his thought is that the conceptual equipment that is operative in perceptual experience generally, whether the experience is such as to yield knowledge or not, is dependent on a world view, in the logical dimension that the metaphor of "foundation" risks leading us to forget. We can capture this part of the picture by saying that the intentionality, the objective purport, of perceptual experience in general—whether potentially knowledge yielding or not—depends, in that logical dimension, on having the world in view, in a sense that goes beyond glimpses of the here and now. It would not be intelligible that the relevant episodes present themselves as glimpses of the here and now apart from their being related to a wider world view in the logical dimension Sellars adds. But the wider world view depends in turn, in the logical dimension that figures in traditional empiricism, on perceptual experience that is capable of yielding knowledge, in the form of glimpses of the here and now. With this mutual dependence, the nontraditional empiricism that Sellars espouses constitutes a picture both of the credentials of empirical knowledge and of the intentionality of empirical thought in general.

This makes it unsurprising that we find Sellars speaking of "the epistemic character, the 'intentionality'," of expressions such as 'thinking of a celestial city'.[7] When he introduces the image of the logical space of reasons, he singles out the episodes or states whose characterizations place them in the space of reasons as episodes or states of *knowing* (*op. cit.*, §36, pp. 298-99 (169)). And of course episodes or states of knowing would have an epistemic character in an etymologically obvious sense. But it would be wrong to conclude that Sellars's concern is narrowly epistemological. In the remark about 'thinking of a celestial city', he makes this clear by showing that he is willing to equate epistemic character with intentionality, and to talk of epistemic character in a case in which there need be no question of knowing. In the remark about 'thinking of a celestial city', 'epistemic' can amount to no more than 'concept involving'.[8] This is the interpretation I announced and promised to vindicate.

I have been urging that Sellars's nontraditional empiricism is not only a picture of the credentials of empirical knowledge, a topic for epistemology in a narrow sense, but also a picture of what is involved in having one's thought directed at the world at all, the topic of reflection about intentionality. This enables me to forestall a possible objection to the proposal that we read Kant, Sellars's model, as a philosopher of intentionality. I do not mean the feeble objection that 'intentionality' is not a Kantian term. 'Intentionality' is a scholastic term, which did not come back into mainstream philosophical currency until (I think) Brentano, but obviously that does not prevent us from supposing that the topic is a Kantian topic. What I have in mind is rather the potentially more challenging objection that Kant's concern is epistemological. As a putative reason for supposing that Kant is not concerned with intentionality, I can neutralize this by saying: certainly, Kant's concern is epistemological—in just the way in which Sellars's is.

---

[7] §7, p. 260 (133). Compare §§24, 25, pp. 283-85 (155-56). Consider also the implication, at §17, p. 274 (146), that looking red is an epistemic as opposed to natural fact about objects. Looking red is not an epistemic fact in the etymologically obvious sense that I mention in the text below.

[8] See *Science and Metaphysics*, p. 23: for purposes of the philosophy of mind, "the intentional is that which belongs to the conceptual order."

Against a "neo-Kantian" reading of Kant, Heidegger says: "The *Critique of Pure Reason* has nothing to do with a 'theory of knowledge'."[9] I think we can make the point Heidegger is trying to make more effectively—certainly we can put it in a form in which it is easier to swallow—by saying, not that epistemology is *no* concern of the first *Critique*, but that it is no more *the* concern of the first *Critique* than it is of "Empiricism and the Philosophy of Mind" or of *Science and Metaphysics*.

III

Early in "Empiricism and the Philosophy of Mind," Sellars diagnoses "the classical concept of a sense datum" as "a mongrel resulting from a crossbreeding of two ideas": first, an idea of non-concept-involving sensory episodes, such as sensations of red, and, second, an idea of noninferential knowings that such-and-such is the case (*op. cit.*, §7, p. 259 (132)). This is a mongrel, a conflation, because attributions of non-concept-involving episodes belong below the line drawn by Sellars's master thought, whereas attributions of knowings belong above it. When Sellars repeats the diagnosis a few pages later, he extracts from it a program for the rest of the lectures: "to examine these two ideas and determine how that which survives criticism in each is properly to be combined with the other" (*op. cit.*, §10, p. 267 (140)). The program, then, is to arrive at an acceptable picture of how the sensory and the conceptual—sensibility and understanding—combine so as to provide for the intentionality of perceptual experience, and (the same function viewed from a different angle) to provide for how perceptual experience figures in the acquisition of a knowledgeable view of the world. We have looked ahead in Sellars's work, so we know that this case of intentionality, the intentionality of perceptual experience, is going to be in one way basic to intentionality in general, though not in a way that involves its being intelligible in advance of the idea of having a world view that goes beyond the immediate deliverances of perception.

The above-the-line element in the mongrel conflation is the idea of noninferential knowings. Sellars mostly focuses on one sensory modality, and considers seeings.[10] But in pursuing his program in connection with this particularization to one sensory modality of the above-the-line element in the mongrel, he expands the topic from seeings to a wider class of experiences, which he initially introduces as ostensible seeings. Seeings are a singled-out subclass of ostensible seeings.[11] Evidently Sellars takes it that for purposes of separating and correctly combining what survives criticism in the ideas that are conflated into the mongrel, what matters is to understand the wider class. The goal is to understand the intentionality of visual experience in general, whether potentially knowledge yielding or not.

---

[9] *Kant and the Problem of Metaphysics*, Richard Taft, trans. (Bloomington: Indiana UP, 1990, fourth edition), p. 11 Heidegger's word is 'Erkenntnistheorie', which might have been translated 'epistemology', see Taft's note, p. 188.

[10] For a self-conscious comment on this, see *Science and Metaphysics*, p. 9. There is a minor complication (nothing turns on it): seeings are not, as such, noninferential knowings or acquirings of knowledge (that was how the above-the-line elements in the mongrel conflation were first introduced), but rather opportunities to know, which may not be taken. Consider how it might be intelligible to say this: "I thought it merely looked to me as if the tie was green, but I now realize that I was seeing it to be green."

Ostensible seeings are experiences in which it looks to their subject as if things are a certain way, and Sellars devotes some effort to elucidating that idea. Centrally important here is the image of an experience as, "so to speak, making an assertion or claim" or as "containing" a claim (*op. cit.*, §16, pp. 271-72 (144)). Sellars introduces this image in an explicitly promissory way, pointing forward to the culmination of "Empiricism and the Philosophy of Mind." There, he vindicates a notion of nonovert conceptual episodes, on the ground that they can be understood by analogical extension from overt conceptual episodes, linguistic acts.[12] Visual experiences "make" or "contain" claims in that they are conceptual episodes, actualizations of conceptual capacities, and as such are to be understood on the model of linguistic performances in which claims are literally made.

This bears some elaboration. I have mentioned the Kantian view that conceptual capacities have their paradigmatic mode of actualization in judgings. We can approach the idea that visual experiences are conceptual episodes, and as such "make" or "contain" claims, through this identification of judging as the paradigmatic kind of conceptual episode. Consider, say, judging that there is a red cube in front of one. There is a conceptual capacity that would be exercised both in making that judgment and in judging that there is a red pyramid in front of one, and another conceptual capacity that would be exercised both in judging that there is a red cube in front of one and in judging that there is a blue cube in front of one. In judging that there is a red cube in front of one, one would be exercising (at least) these two capacities together. What does 'together' mean here? Not just that one would be exercising the two capacities in a single act of judgment; that would not distinguish judging that there is a red cube in front of one from judging, say, that there is a red pyramid and a blue cube in front of one. In a judgment that there is a red cube in front of one, the two conceptual capacities I have singled out would have to be exercised with a specific mode of togetherness: a togetherness that is a counterpart to the "logical" or semantical togetherness of the words 'red' and 'cube' in the verbal expression of the judgment, 'There is a red cube in front of me'. Here we see the point of the idea that nonovert conceptual episodes are to be understood on analogy with linguistic acts; it affords a way to make a distinction that we need to make.[13]

---

[11] For seeings as veridical members of a class of ostensible seeings, see "Empiricism and the Philosophy of Mind," §7, pp. 260-61 (133-34). This points toward the discussion of 'looks' statements in §III, pp. 267-77 (140-49). I shall comment in my third lecture on the idea, which is implicit in at least the first version of "Empiricism and the Philosophy of Mind," that veridicality is all it takes for an ostensible seeing to be a seeing

[12] The first phase of the myth of Jones; for the application to perceptual experience, see §60, p. 321 (190).

The conceptual episodes Sellars is concerned with, when he speaks of visual experiences as "containing" claims, are not as such cases of judging. Even if one does judge that things are as they look, having them look that way to one is not the same as judging that they are that way. In some cases, perhaps, one does judge that things are a certain way when they look that way—acquiring the belief that they are that way by freely making up one's mind that they are that way. But more typically, perceptual belief acquisition is not a matter of judging, of actively exercising control over one's cognitive life, at all. Unless there are grounds for suspicion, such as odd lighting conditions, having it look to one as if things are a certain way—ostensibly seeing things to be that way—becomes accepting that things are that way by a sort of default, involving no exercise of the freedom that figures in a Kantian conception of judgment.

So there is a disconnection between perceptual experience and judging. But even so, we can exploit the apparatus of the conception of judging I have sketched in order to vindicate Sellars's image of experiences as "containing" claims. A free, responsible exercise of certain conceptual capacities, including at least the two I mentioned, with a suitable mode of togetherness would be judging that there is a red cube in front of one. Now we can say that in an ostensible seeing that there is a red cube in front of one—an experience in which it looks to one as if there is a red cube in front of one—the *same* conceptual capacities would be actualized with the *same* mode of togetherness. This cashes out the idea that an experience so described "contains" a claim, whose content is just what one would be judging in the corresponding judgment. But this actualization of the relevant conceptual capacities, unlike the one that would be involved in the corresponding judgment, would be involuntary; that is why I say 'actualization' rather than 'exercise'.

---

[13] I mean this quick sketch of a conception of judging, as the joint exercise of different conceptual capacities, to recall Gareth Evans's discussion of "the Generality Constraint": *The Varieties of Reference* (New York: Oxford, 1982), pp. 100-05. Evans's discussion has its roots in P T. Geach's account of judging on analogy with saying, in *Mental Acts* (London· Routledge, 1957). Geach's analogical account of judging is roughly contemporary with Sellars's in "Empiricism and the Philosophy of Mind"; independent of it; and (I would argue) more satisfactory, in being free of the scientistic baggage with which Sellars encumbers his version. But I shall not be considering the scientistic details of Sellars's version in these lectures (though Sellars's scientism will matter in other contexts).

This idea of conceptual capacities being involuntarily actualized in perceptual experience partly captures the point of a striking remark Sellars makes about the way an experience "contains" a claim; he says that the claim "is, so to speak, evoked or wrung from the perceiver by the object perceived" (*op. cit.*, §16 bis, p. 272 (144)). When Sellars says this, he is talking about experiences of seeing, but the point he is making surely applies also to members of the wider class, ostensible seeings, even the ones that are not seeings. Ostensible seeings are experiences that, as conceptual episodes, "contain" claims, but in a special way that differentiates them from conceptual episodes of other kinds. They "contain" their claims as ostensibly *necessitated* by an object ostensibly seen. In *Science and Metaphysics*, Sellars puts the same point by saying that, if one says it looks to so-and-so as though there were a red and rectangular physical object in front of him, one is attributing to so-and-so (who may, of course, be oneself) a conceptual representation, of a particular kind, that there is a red and rectangular physical object in front of him; and the kind is "that kind of conceptual representation which is being under the visual *impression* that...there is (or of there being) a red and rectangular physical object in front of one" (*op. cit.*, p. 14; my emphasis). In the language of "Empiricism and the Philosophy of Mind," this is to say that ostensible seeings "contain" their claims in a distinctive way, one that distinguishes them from other conceptual episodes; they "contain" their claims as ostensibly visually *imposed* or *impressed* on their subject.[14]

---

[14] I have, I think charitably, discounted "evoked or wrung *from the perceiver*" in the formulation Sellars uses in "Empiricism and the Philosophy of Mind." A claim evoked from a perceiver would surely be a claim that the perceiver makes. But it seems wrong to imply that a perceiver makes the claim his experience "contains"—wrong even before we widen the focus from seeings to ostensible seeings. Whether an ostensible seeing is a seeing turns on whether its "contained" claim is true, and that is a separate question from whether its subject makes (endorses) the claim. (See footnote 10 above.) So even a seeing, let alone a merely ostensible seeing, does not necessarily "contain" a claim made by its subject. Where I have marked an omission in my citation of the parallel remark from *Science and Metaphysics*, Sellars glosses "being under the visual impression that" with "(visually taking it to be the case that)," and this seems wrong in the same way. We can correct Sellars on this without posing a threat to something he wants to insist on: that one gets to have conceptual episodes (representations) of the relevant kind occur in one's life at all only by acquiring the capacity to make the claims they "contain."

So it is not simply that conceptual episodes of the relevant kind consist in actualizations of conceptual capacities that are involuntary. (We have that also with other kinds of conceptual episode; for instance, when one is, as we say, struck by a thought.) In visual experiences, conceptual capacities are actualized with suitable modes of togetherness; this is how we cash out the idea that the episodes "contain" claims. But they are actualized with an involuntariness of a specific kind; in a visual experience an ostensibly seen object ostensibly impresses itself visually on the subject. Presumably parallel things are to be said about other sensory modalities.[15]

IV

I have been considering what survives criticism, from the above-the-line element of the mongrel conflation, in a conception of visual experience that would be acceptable by Sellars's lights. His program for "Empiricism and the Philosophy of Mind" presupposes in addition that something survives criticism from the below-the-line element in the mongrel conflation, and needs to be acceptably combined with conceptual episodes of the distinctive kind I have been discussing, in a total picture of visual experience. In the course of executing his program, Sellars says it is "clear" that there is more to visual experience than conceptual episodes of that distinctive kind; and, specifically, that a full picture must also include non-concept-involving episodes of the kind exemplified, in the original description of the mongrel conflation, by sensations of red.[16] But why is this supposed to be clear?

The question is especially pressing when we realize how much goes into what we already have above Sellars's line—how much goes into the idea of a conceptual episode of the relevant kind. Even after we have said not just that a visual experience "contains" a claim but also that the "contained" claim is, so to speak, "evoked" by an ostensibly seen object, Sellars still says it is clear that we need to add something about visual episodes of a nonconceptual kind. In his view, conceptual episodes of the relevant kind are already, as the conceptual episodes they are, cases of *being under the visual impression* that such-and-such is the case. It is not that as conceptual episodes they are phenomenologically colorless, so that they would need to be associated with visual sensations in order that some complex composed of these conceptual episodes and the associated visual sensations can be recognizably visual. These conceptual episodes are already, as the conceptual episodes they are, shapings of visual consciousness.[17] If we need a below-the-line element in our picture, it is not in order to ensure that the picture depicts states or episodes of visual consciousness.

---

[15] Compare the conception of experience I recommended in *Mind and World*, where I wrote of states or episodes in which conceptual capacities are operative in sensibility. I think such a formulation simply captures, in explicitly Kantian language, the way Sellars shows us how to conceive perceptual experience—at any rate what he sees as the above-the-line element in the total truth about perceptual experience. In *Mind and World* (for example, pp. 140-41) I focused on the below-the-line role that Sellars credits to sensibility, and missed the fact that he has an above-the-line conception of perceptual impressions that matches the conception I was recommending.

[16] "Empiricism and the Philosophy of Mind," §16 bis, p. 272 (144-45). Compare the use of 'of course' at §22, p. 279 (151), and again at §45, p. 305 (175).

So why does Sellars think our total account of visual experience needs to include visual sensations as well? About the presence of the corresponding element in the mongrel conflation, he says:

> [This] idea clearly arises in the attempt to explain the facts of sense perception in scientific style. How does it happen that people can have the experience which they describe by saying "It is as though I were seeing a red and triangular physical object" when either there is no physical object there at all, or, if there is, it is neither red nor triangular? The explanation, roughly, posits that in every case in which a person has an experience of this kind, whether veridical or not, he has what is called a 'sensation' or 'impression' 'of a red triangle' (*op. cit.*, §7, p. 259 (132-33)).

And in the view that emerges as his own, in the course of "Empiricism and the Philosophy of Mind," it is not in this explanatory motivation for its below-the-line element that the mongrel conflation goes wrong, but just in the way it conflates the below-the-line element so motivated with the above-the-line element, episodes that would have to be actualizations of conceptual capacities. Sensations figure in the picture, at least initially, as posited on the ground that they are needed for an explanatory purpose.[18]

What explanatory purpose? In "Empiricism and the Philosophy of Mind," the envisaged explanation is, as we have seen, "in scientific style." And the question that the explanation is to answer seems to be this, to put it in terms that become available during the execution of Sellars's program: How is it that the *same* claim would be "contained" in, say, each member of a trio of possible experiences of which one is a case of seeing that there is a red and triangular physical object in front of one, one is a case in which something in front of one looks red and triangular although it is not, and one is a case in which it looks to one as if there is something red and triangular in front of one although there is nothing there at all?[19]

---

[17] Contrast, for instance, Robert B. Brandom, *Making It Explicit: Reasoning, Representing, and Discursive Commitment* (Cambridge: Harvard, 1994). In his chapter 4, Brandom undertakes to give an account of observational claims and observational knowledge while sedulously avoiding any mention of sensory consciousness. Brandom here diverges from something that is quite central to Sellars's thinking.

[18] See §§21-22, pp. 277-80 (149-52); a programmatic passage, and he returns to its program in §45, pp. 305-07 (175-77), and executes it in the rest of the lectures. Much of the work needed is in qualifying the idea of sensations as posited, in order to make room for immediate self-attribution of sensations. This is why I say 'at least initially'.

If what we still need to ask for is an explanation of the *sameness* of claims "contained" in such a trio of experiences, the request for an explanation apparently assumes that we are already entitled to think of experiences as "containing" some claims or other, independently of the explanation we are asking for. After all, we might say, the idea that experiences "contain" claims is already accounted for in the part of the story that belongs above the line; we have already entitled ourselves to it by talking about actualizations of conceptual capacities, before questions arise about a below-the-line element in the total story. But in that case, it is not clear why we should suppose that our explanatory need can be met only by finding a sameness at the level of visual *sensations*—items in consciousness—between the members of such a trio, as opposed to a sameness at the level of, say, patterns of light impinging on retinas. When he elaborates "the attempt to explain the facts of sense perception in scientific style" by positing sensations, Sellars himself says: "The core idea is that the *proximate cause* of such a sensation is only for the most part brought about by the presence in the neighborhood of the perceiver of a red and triangular physical object" (*op. cit.*, §7, pp. 259-60 (133); emphasis altered). In this remark, Sellars is suggesting that we should expect to find a sameness between seeings and ostensible seeings that "contain" the same claims, at a level that he here specifies as that of *proximate causes* of sensations—for instance, at the level of retinal images. But then why not suppose a sameness at this level will do the explanatory work for which Sellars thinks we need to appeal to sensations? Conceptual episodes of the relevant kind are triggered by impacts from the environment on a perceiver's sensory equipment.

---

[19] See §45, p. 305 (175) for a formulation on these lines. Sellars notes (§22, p. 279 (151)) that it is not strictly accurate to say that the same claim is "contained" in each member of such a trio: the claim "contained" in each of the first two is referential, whereas the claim "contained" in the third is not. This will start to be significant in my second and third lectures, but it does not matter here, any more than it does at the point where Sellars acknowledges it.

If the impacts are suitably similar, there is nothing puzzling about a similarity between the conceptual episodes they trigger. And it is not clear why it should seem necessary to describe these suitably similar impacts in terms of nonconceptual impingements *on consciousness* (sensations), as opposed to saying that consciousness comes into play only with conceptual episodes, triggered by nonmentalistically described impacts on sensory equipment. It seems that what Sellars here introduces as proximate causes of sensations can themselves meet the explanatory need, conceived as he seems to conceive it in "Empiricism and the Philosophy of Mind." The sensations look like idle wheels.

In *Science and Metaphysics,* Sellars explicitly confronts an objection on these lines (*op. cit.,* p. 18); and he responds in a way that changes the picture rather radically from the one he seemed to be giving in "Empiricism and the Philosophy of Mind." First, he no longer formulates the explanation-seeking question in terms of the sameness of the claims "contained" in different possible experiences—as if we could anyway help ourselves to the idea that experiences "contain" claims at all. The explanation-seeking question now is: How is it that sensory relatedness to the environment takes the form of conceptual episodes, episodes that, in the terminology of "Empiricism and the Philosophy of Mind," "contain" claims, at all?[20] Second, the explanatory need that sensations are supposed to satisfy is not a need for scientific understanding, as it seemed to be in "Empiricism and the Philosophy of Mind"; rather, it is *transcendental.*[21] I think these are two ways of putting the same thought: the reason Sellars thinks our complete account of visual experience must include visual sensations—nonconceptual visual episodes—is that he thinks this is the only way we can find it intelligible that there should so much as be the conceptual shaping of sensory consciousness that constitutes the above-the-line element in his account of visual experience.

---

[20] See *op. cit.,* p. 18, where, in urging that the explanatory question is not specially about nonveridical experiences, Sellars writes: "even in normal [veridical] cases there is the genuine question, 'Why does the perceiver *conceptually represent* a red (blue, etc.) rectangular (circular, etc.) object in the presence of an object having these qualities?' "

[21] See *op. cit.,* p. 9. Sellars says that manifolds of sensation are "postulated on general epistemological or, as Kant would say, transcendental grounds." I think this equation, in the context of the understanding I am offering of the sense-impression inference, reinforces the impression given by his willingness to equate 'intentionality' with 'epistemic character' in "Empiricism and the Philosophy of Mind"; he does not conceive epistemology narrowly.

'Transcendental' figures here in a recognizably Kantian sense. The explanation Sellars envisages is transcendental because it is needed, he thinks, in order to vindicate the legitimacy of the apparatus—the talk of experiences as actualizations of conceptual capacities, which as such "contain" claims, but in a distinctively sensory way—in terms of which we enable ourselves to conceive experiences as ostensibly *of objects* at all.[22] Sellars thinks his picture, with sensations playing such a transcendental role, just is the picture Kant would have given if he had been fully clear about the drift of his own thinking.

On this reading of the "sense-impression inference" as it figures in *Science and Metaphysics* (for the phrase, see *op. cit.*, p. 17), visual sensations or sense impressions are not simply an extra part of the truth about visual experiences, over and above the part that deals with the distinctive way in which visual experiences "contain" claims. That is how it might have seemed from "Empiricism and the Philosophy of Mind." But in the view Sellars urges in *Science and Metaphysics*, it is not that visual experiences "contain" claims in their distinctive way, and then there is a simply additional fact about them, namely, that they involve visual sensations. The reason we have to acknowledge the "additional" fact, in Sellars's view, is that only so can we be entitled to have spoken as we did when we gave our above-the-line characterization of visual experiences—when we spoke of visual experiences as "containing" claims, and so having objective purport, in the distinctive way they do.

Sellars's sense-impression inference is a piece of transcendental philosophy, in the following sense: it is directed toward showing our entitlement to conceive subjective occurrences as possessing objective purport. Notice that that description of transcendental philosophy implies nothing in particular about the nature of the activity. There is a temptation to suppose transcendental philosophy would have to be done at a standpoint external to that of the conceptual goings-on whose objective purport is to be vindicated—a standpoint at which one could contemplate the relation between those conceptual goings-on and their subject matter from sideways on. Sellars's move fits this conception; he undertakes to vindicate the objective purport of conceptual occurrences from outside the conceptual order. I shall be taking issue with this conception of transcendental philosophy. It is important to see that this is not to take issue with the very idea of transcendental philosophy.[23]

---

[22] See, for example, *Critique of Pure Reason*, Norman Kemp Smith, trans. (London: Macmillan, 1929), A11-12/B25: "I entitle *transcendental* all knowledge which is occupied not so much with objects as with the mode of our knowledge of objects in so far as this mode of knowledge is to be possible *a priori*."

V

When Sellars vindicates the idea of inner episodes, at the culmination of "Empiricism and the Philosophy of Mind," he does so in two phases: first for conceptual episodes and then for nonconceptual episodes. Each phase has two stages, and the structure is parallel in each phase. First, there is an account of how concepts of episodes of the relevant kind could have been introduced in the context of a theory; at this stage, the episodes are envisaged as attributable, to others or oneself, only inferentially, in a way mediated by the theory. But then, second, there is an account of how a noninferential self-attributing ("reporting") employment of the relevant conceptual apparatus could have been introduced, by training people in such a way as to leave them immediately disposed to make self-attributions—'immediately' in the sense that they do not need to advert to the evidence that the theory provides for—on occasions when, according to the theory, those attributions are correct. By the end of "Empiricism and the Philosophy of Mind," conceptual episodes, including those which "contain" claims in the distinctive way in which visual experiences do, and nonconceptual sensory (and in particular visual) episodes—impressions or sensations—are on a level, in respect of being available for noninferential self-attribution.[24]

In *Science and Metaphysics*, Sellars modifies this picture in a way that belongs, I think, with the fact that he now explicitly sees the sense-impression inference as transcendentally driven. He suggests that the visual impressions or sensations that the sense-impression inference requires us to posit are states of consciousness, but not objects of consciousness; or—the same thought differently expressed—that these impressions or sensations are states of consciousness that are not apperceived, where "apperception" can be explained as "noninferential self-knowledge."[25] He is suggesting, then, that the transcendentally posited visual impressions or sensations figure in visual consciousness in a way that does not amount to their achieving the immediate or noninferential attributability to oneself that he works to secure at the culmination of "Empiricism and the Philosophy of Mind."

---

[23] The idea that transcendental philosophy would have to be done from a special standpoint is implicit in Richard Rorty, *Philosophy and the Mirror of Nature* (Princeton: University Press, 1979), p. 293, where Rorty writes of the "demand...for some transcendental standpoint outside our present set of representations from which we can inspect the relations between those representations and their object." Kant distinguishes "transcendental" and "transcendent" (see, for example, A296/B352-53). In Rorty's phrase, 'transcendental' could be replaced by 'transcendent'. Not that that shows Rorty to be misusing 'transcendental'; he is suggesting that transcendental philosophy requires a transcendent standpoint. That is what I think we should dispute. When I wrote disparagingly about Kant's "transcendental story" in *Mind and World* (pp. 41-43, 95-98), I was acquiescing, in a way I now regret, in a reading of parts of Kant's transcendental activity that fits Rorty's phrasing. (I would still disparage the philosophy such a reading finds in Kant.)

[24] See §59, p 320 (189) for conceptual episodes, and §62, p. 327 (194-95) for nonconceptual episodes.

Now it is hard to see how, on Sellarsian or indeed any principles, there could be a class of items in consciousness whose members were permanently and constitutionally incapable of being apperceived, incapable of being directly available for self-attribution. Sellars cannot mean to be suggesting that the visual impressions or sensations that, according to his transcendental sense-impression inference, must figure in episodes of "outer sense" are, simply as the visual impressions or sensations they are, incapable of being objects of consciousness. I think his thought must rather be on the following lines. The visual impressions or sensations in question are not apperceived *when they are playing their transcendental role.* That is not to say that they are not *apperceivable.* It is just to say that if they do get to be apperceived—if they do become objects for consciousness—they can no longer be playing their transcendental role, that of enabling episodes of "outer sense," episodes that "contain" claims about the environment. One can focus one's attention on the manifold of "sheer receptivity" that was, a moment before, enabling one's attention to be directed toward the ostensibly seen environment. But in doing so—in bringing it within the scope of one's apperception—one ensures that it ceases to perform that function.

I think this thought strengthens Sellars's position by immunizing it against a certain objection. If it were right to endorse Sellars's sense-impression inference, it would be a good idea to construe its conclusion in this way. Considering only "Empiricism and the Philosophy of Mind," I used to think one could complain that the below-the-line items that figure in Sellars's picture of visual experience would be opaque; not something through which the environment could intelligibly be revealed to us, but at best something on the basis of which, if we knew enough about how features of the environment cause these affections of our sensory capacities, we could infer conclusions about the environment.[26] In "Empiricism and the Philosophy of Mind," the only way Sellars considers for something to be in consciousness is for it to be an object of consciousness. This leaves it seeming that the sensations that are part of Sellars's picture of perceptual experience would have to be objects of consciousness, on pain of not figuring in consciousness at all. If that were not what Sellars wants, why does he need to work at securing that they can be objects of consciousness, as he does at the culmination of "Empiricism and the Philosophy of Mind"? And if we try to make out that they are objects of consciousness, it does seem that they would engross the attention, and prevent it from fixing itself—except indirectly, through inference—on environmental objects, although it was perception of environmental objects that we were supposed to be trying to make intelligible.

---

[25] *Science and Metaphysics*, pp. 10, 11 For the gloss on "apperception," see p. 72

[26] See *Mind and World*, p. 145.

But the picture Sellars gives in *Science and Metaphysics* is immune to any such objection. Now he can concede that, if a manifold of visual sensations is figuring as an object for its subject's consciousness, the subject's attention can no longer pass through it to features of the environment directly, but at best inferentially. But this leaves unthreatened the idea that such a manifold is transcendentally required for perceptual awareness of the environment. Sellars is now equipped to say that the transcendentally required manifold, when it is doing its transcendental job, figures in consciousness not as an object, in which case it would indeed prevent the free passage to the environment of the subject's attention, but precisely as that through which the subject's attention is directed without hindrance to features of the ostensibly seen environment. The idea is that attention, which involves apperception, can be directed either at the ostensibly seen environment or at the visual sensations that were enabling the environment to be ostensibly seen, but not both; if the attention is directed at the sensations, they can no longer be enabling the ostensible seeing of environmental objects.

This complication, however, does not undermine the fact that a Sellarsian account of the noninferential self-attribution of visual sensations—when it does occur, as it surely can—will have the structure established in "Empiricism and the Philosophy of Mind." The concepts under which visual sensations are apperceived when they are—which is not when they are enabling episodes of "outer sense"—will be concepts whose original home is a transcendentally required theory concerning how manifolds of sensations enable episodes of "outer sense." When visual sensations become objects for consciousness, it will be under concepts whose original function is to connect these episodes, in a theory-mediated way, with the claim-"containing" character of visual experiences. Some people think we can vindicate a role for sensations in the total picture of visual experience on the basis of a theoretically innocent introspection. The idea is that sensational properties are introspectively available in any case, whatever we say about what figures, in this conception, as the subsequent question how, if at all, the sensational properties of experiences relate to their claim-"containing" character. I have ignored this conception, because it makes no contact with Sellars's thinking.[27]

It is important not to be misled by the fact that Sellars uses the word 'impression' both in the phrase 'being under the visual impression that... ', which characterizes a kind of conceptual episode, and for the below-the-line element in his picture of perceptual experience.[28] Conceptual episodes that belong to the kind, being under the visual impression that... , are, simply as conceptual episodes, available for apperception when they occur.[29] Impressions in the other sense, in contrast, can be apperceived only when they are not serving as the below-the-line element in the total truth about some perceptual experiences.

---

[27] See, for example, Christopher Peacocke, *Sense and Content* (New York: Oxford, 1983), chapter 1. Peacocke says the sensational properties of visual experience are arrayed in a two-dimensional visual field, and he does not suggest that 'two-dimensional' here means anything different from what it might mean in describing, say, a surface in the environment. By Sellars's lights this is naive, for Sellars, the spatiality of the arrangement of visual sensations is not the spatiality of "outer" configurations, but something that needs to be understood by analogical extension from it, in a sophisticated exercise of concept formation.

Even on their own terms, I think Peacocke's phenomenological arguments are unconvincing, but I shall not argue this here. Perhaps the thinness of the supposedly independent phenomenological considerations reveals that Peacocke's conception of what a supposedly innocent introspection would yield is really controlled by an implicit acceptance of something like Sellars's transcendental thought.

There is another putative ground for supposing that visual experiences must have a sensational aspect, equally non-Sellarsian, which I shall also not discuss in these lectures. This is the thought that there must be a two-dimensional sensational array to serve as a vehicle for the representational content of a visual experience, somewhat as an arrangement of pigment on a surface is a vehicle for the representational content of a picture.

[28] See *Science and Metaphysics*, p. 19, where Sellars distinguishes "an impression of a red rectangle" from "an impression of a man lurking in the corner." The latter would, as he says, be "a conceptual state" (or episode); one of the kind identified at p. 14, the kind "being under the visual impression that...."

[29] This is not to say that they are actually apperceived. The 'I think' of apperception must be *able* to accompany all my representations (*Critique of Pure Reason*, B131), which is not to say that it actually accompanies them. But conceptual representations are available for apperception in a way that differs from that in which, in a plausibly Sellarsian picture, perceptual sensations are; apperceiving the latter would require equipping oneself with something new, a conceptual representation involving concepts whose primary home is the transcendental theory of how conceptual representations of outer reality are guided by manifolds of "sheer receptivity."

The fact that the same word naturally acquires these two uses is perhaps suggestive about the shape of Sellars's picture. When a conceptual episode is apperceived as belonging to the kind, being under the visual impression that..., what is apperceptively available, according to Sellars's picture, is *that* the flow of one's conceptual representations, of the sort involved in normal perceptual activity, is being guided into "containing" the relevant claim by the flow of one's impressions in the below-the-line sense. Perhaps this is why 'being under the visual impression that...' is an appropriate specification of the kind to which the conceptual representation apperceivably belongs. But apperception does not embrace the specifics of *how* this guidance is effected; if the formerly guiding items get to be apperceived, they can no longer be performing their guiding function.[30]

VI

Sellars thinks this picture is essentially the one Kant is aiming at, although he has to acknowledge that it is not to be found adequately set out on Kant's pages. For one thing, Kant "tends to restrict the term 'consciousness' to apperceiving and the apperceived as such," which makes it difficult to find in him the idea that impressions or sensations can figure in consciousness without being apperceived (*ibid.*, p. 11). A more substantial problem is that Sellars has to find Kant seriously confused in his thesis that space is the form of outer sense.[31] One is bound to wonder whether Sellars has Kant wrong; and, since Sellars's reading of Kant is, perfectly properly, shaped by Sellars's own conviction about how we should conceive perceptual experience, that is inextricably bound up with wondering whether Sellars is mistaken in thinking that sound philosophy requires impressions or sensations to be credited with the role he attributes to them. These are my questions for my next lecture.

---

[30] For the image of guiding, see *Science and Metaphysics*, p. 16.

[31] See *ibid.*, p. 8: "the idea that Space is the form of outer sense is incoherent."

LECTURE II: THE LOGICAL FORM OF AN INTUITION

In Lecture I, I elicited from Sellars a picture of the intentionality of perceptual experience—visual experience, to stay with the case Sellars mostly concentrates on.

His picture has elements both above and below a line that importantly shapes his thinking. The line separates characterizations of occurrences in people's lives that need to be understood in terms of the actualization of conceptual capacities from characterizations that do not need to be understood in those terms.

Above the line in a Sellarsian picture of a visual experience, there is a conceptual episode of a distinctive kind. Just by virtue of being a conceptual episode, such an episode "contains" a claim about the environment. But episodes of this kind are differentiated from conceptual episodes of other kinds in that they "contain" their claims in a distinctive way: as ostensibly required from or impressed on their subject by an ostensibly seen object.

Below the line in a Sellarsian picture of a visual experience, there is a complex or manifold of visual sensations, that is, non-concept-involving visual episodes or states. Why does Sellars think the picture has to include this element as well as conceptual episodes of the relevant kind? Not to ensure that the picture respects phenomenological facts—as if there would be nothing sensory, let alone visual, about the episodes that are in view before we advert to this below-the-line element. On the contrary, the above-the-line episodes that figure in Sellars's picture of visual experience are, as conceptual episodes of their special kind, already conceived as conceptual shapings of sensory, and in particular visual, consciousness. Sellars's thought is rather this: it is for transcendental reasons that we need to acknowledge the below-the-line element in the picture. The idea is that we are entitled to talk of conceptual episodes in which claims are ostensibly visually impressed on subjects—the above-the-line element in the picture—only because we can see the flow of such conceptual representations as guided by manifolds of sensations, non-concept-involving episodes or states in sensory, and specifically visual, consciousness.

Sellars takes this picture to be fundamentally Kantian, although he complains that Kant failed to make the necessary distinction clear even to himself. As I suggested at the end of Lecture I, this invites us to consider a different exegetical possibility. Perhaps the idea that perception involves a flow of conceptual representations guided by manifolds of "sheer receptivity" is not Kantian at all. I am going to urge that that is indeed so; the idea is foisted on Kant by Sellars, even with his eyes open to the price, which is that he needs to accuse Kant of confusion. Sellars is willing to pay the price, because he is convinced that the idea is required for a satisfactory execution of Kant's project. In Lectures II and III, I am going to urge that Sellars is wrong about that, too. I want to suggest that, so far from helping to make us comfortable with the intentionality of perception, and thereby contributing toward making us comfortable with intentionality in general, the below-the-line element in Sellars's picture actually stands in the way of a useful conception of how perception and thought are directed toward objects—one we can find, at least in germ, in Kant, once we discard Sellars's interpretation of the transcendental role Kant credits to sensibility.

I

Sellars is firm in his conviction that what Kant usually calls *intuitions* are representations of individuals that already involve the understanding, the faculty associated with concepts. He suggests that an intuition on this interpretation of the term should be taken to represent an individual as a *this-such*.[1] I think this is very helpful, and I am going to exploit it. In an intuition on this interpretation of the term, sensibility and understanding are both involved. On this interpretation, we might describe intuitions as shapings of sensory consciousness by the understanding—to echo the wording I used in connection with Sellars's conception of the conceptual episodes, as such "containing" claims, that figure above the line in the picture of perceptual experience he gives in "Empiricism and the Philosophy of Mind."

But as I have said, Sellars is convinced that Kant also needs to speak about sensibility in a way that belongs below his line, as the talk of sensory consciousness with which we can gloss this first notion of intuition does not, because in intuitions, on this first interpretation, sensory consciousness is already shaped by the faculty of concepts. Sellars thinks the transcendental role that Kant needs sensibility to play consists in its supplying manifolds of sensory items that are not shaped by the understanding, to guide the flow of conceptual representations in perception. So Sellars thinks Kant needs the word 'intuition'—his most general term for sensibility in operation—to apply also to occurrences in which the understanding, the faculty of concepts, is not involved: "We seem...to be led to a distinction between intuitions which do and intuitions which do not involve something over and above sheer receptivity" (*ibid.*, p. 4; also p. 7 (§17)).

---

[1] See *Science and Metaphysics*, p. 3: "On this model, which I take to be, on the whole, the correct interpretation, intuitions would be representations of *thises* and would be conceptual in that peculiar way in which to represent something as a *this* is conceptual." This is elaborated at pp 4-7, where Sellars suggests that representing something as a *this* is representing it as a *this-such*.

Sellars sees it as a failure on Kant's part that he does not distinguish these two interpretations of 'intuition'. And Sellars sees this failure as an implicit counterpart to the mongrel conflation that, in "Empiricism and the Philosophy of Mind," he finds in the classical concept of a sense datum:

> Kant's use of the term 'intuition', in connection with human knowledge, blurs the distinction between a special sub-class of *conceptual* representations of individuals which, though in some sense a function of receptivity, belong to a framework which is in no sense prior to but essentially includes general concepts, and a radically different kind of representation of an individual which belongs to sheer receptivity and is in no sense conceptual.[2]

This is just like the mongrel conflation: a failure to separate items that belong above and below Sellars's line.

The fact that Sellars finds this implicit conflation in Kant helps to account for some features of the above-the-line notion of intuitions that he finds in Kant. In Sellars's reading, Kant's above-the-line notion of intuitions is distorted, in some of its applications, by the implicit conflation with a below-the-line notion.

An intuition on the above-the-line interpretation represents its object as a *this-such*. For instance, a visual intuition might represent its object as *this cube*. Now what does the word 'cube' contribute to such a specification of the content of an intuition? According to Sellars, Kant thinks it can be something prior to the concept of a cube, as that concept might figure in a judgment that something is a cube. In Sellars's reading, Kant thinks the concept of a cube—the concept that figures predicatively in such a judgment—is derived, by an analytic activity of the understanding, from something that is not yet that concept, figuring in intuitions that, even though they do not involve that concept, are nevertheless enabled to represent their object as *this cube* by a synthetic operation of the understanding functioning in the guise of the productive imagination (*ibid.*, pp. 4-7).

---

[2] *Science and Metaphysics*, p 7, compare "Empiricism and the Philosophy of Mind," §7, p 259 (132).

On this view, *cube* in a representation of an object as *this cube* can be prior to *cube* in a judgment that something *is a cube*. As Sellars remarks, this suggestion of a priority is "puzzling" (*ibid.*, p. 5). But it might seem to make sense in the context of the counterpart Sellars finds in Kant to the mongrel conflation of "Empiricism and the Philosophy of Mind." The idea Sellars finds in Kant is that some intuitions are only protoconceptual. This can be seen as a response to the pressure—ex hypothesi not properly understood by Sellars's Kant, who is implicitly entangled in the mongrel conflation—to have the mongrel notion of an intuition provide for an idea that can in fact become clear only when the conflation is unmasked: only as the idea of an episode that is not conceptual at all.

Sellars says: "Kant's thesis...requires the existence of completely determinate 'basic' perceptual this-suches" (*ibid.*, p. 7). This fits the view that Kant thinks concepts proper, at least at a basic level, are derived by abstraction from representations of *this-suches* that are only protoconceptual. Taking Kant's thinking to be distorted by an implicit commitment to the mongrel conflation fits with finding in Kant, as Sellars does, a conception of how the most basic empirical concepts are formed that is in a certain sense abstractionist.[3]

II

On Sellars's view of the requirements for a properly Kantian position, the "Transcendental Aesthetic" should have dealt with forms exemplified in manifolds of intuition on the second interpretation of the term—manifolds of sensory impressions that are prior to any operations of the understanding, and that transcendentally subserve intuition on the first interpretation of the term, according to which intuitions involve the understanding as well as sensibility. Kant's topic in the "Aesthetic," insofar as the "Aesthetic" bears on "outer sense," should have been a form exemplified in the manifolds of "sheer receptivity" that transcendentally enable outwardly directed episodes in which the understanding is also operative. Kant says the form of outer sense is space. If his topic had been what Sellars thinks it should have been, Kant would have had to mean that space informs the below-the-line element in the transcendental picture of outer awareness as a shaping of sensory consciousness by the understanding. But—as Sellars complains—space, as Kant considers it in the "Aesthetic," is a form of outer intuitability on the first interpretation, a form of availability to intuition on the interpretation according to which intuition already involves the understanding. It is a form in accordance with which the *this-suches* that are objects of outer intuition on the first interpretation of the term, episodes of sensory consciousness shaped by the understanding, are given to the subjects of those episodes. That is, what space informs is above the line, and the "Aesthetic" fails to discuss what it should have discussed, a form that does its informing work below the line.[4]

---

[3] The abstractionism is of a peculiar kind. Elsewhere (for example, "Phenomenalism," in *Science, Perception and Reality*, p. 90) Sellars credits Kant with seeing that concepts cannot be abstractively derived from *sensation*, in a process we would have to picture in terms of content being transferred into the intellect "as Jack Horner transferred the plum" (compare *Science and Metaphysics*, p. 20). In the position he attributes to Kant in *Science and Metaphysics*, it is not sensation from which concepts are abstractively derived, but intuitions conceived as already shaped by the understanding. But the position still violates a basic Sellarsian conviction, namely, that the capacity to experience things as thus-and-so should be seen as coeval with the capacity to judge that they are thus-and-so.

Largely below the surface in *Science and Metaphysics* is a detailed picture of how the productive imagination generates intuitions out of (strictly) sensory material, which helps account for the view of concept formation Sellars attributes to Kant. See Sellars's "The Role of the Imagination in Kant's Theory of Experience," in Henry W. Johnstone, Jr., ed., *Categories. A Colloquium* (State College: Penn State UP, 1978). I cannot go into this here.

In making this complaint, Sellars ignores a certain reading of how the "Aesthetic" fits in the overall scheme of the *Critique*. On this reading, we are supposed to account for the outwardness of outer sense by invoking space as an autonomous form of sensibility, intelligible independently of any involvement on the part of the understanding. When Kant then brings the understanding into play, in the "Transcendental Analytic," the outwardness for which, on this reading, the "Aesthetic" has already provided takes on a new form, as directedness toward determinate *objects*. On this reading, space as the "Aesthetic" considers it would after all do its informing work below something corresponding to Sellars's line, with operations of the understanding above the line.

Sellars's complaint leaves no room for this reading. The complaint implies that as he reads Kant, the idea that space informs an outward directedness of subjectivity cannot be understood independently of the idea that objects are available to outer intuition on the first interpretation of the term, according to which intuitions are shapings of sensibility by the understanding. That is, space as the "Aesthetic" considers it, under the title of the form of outer sense, is not meant to be fully intelligible until we have the "Analytic" as well as the "Aesthetic." For Sellars, space as an autonomous form of "sense as such"—what Kant, he thinks, should have concerned himself with—would have to be, not an already outer matrix or arena waiting, as it were, to be determinately populated with objects, but a form of inner states or episodes. Our comprehension of it would have to be constructed by analogical extension from our comprehension of space as the outer matrix in which intuitions on the first interpretation, shapings of sensibility by the understanding, locate objects.[5]

---

[4] See *Science and Metaphysics*, pp. 8, 28-30.

I think Sellars here shows a fine understanding of how Kant intends the thesis that space is the form of outer sense, and a fine understanding of the possibilities for making sense of spatiality as the matrix in which outer objects are given to us. The reading Sellars ignores does not fit Kant, and it does not make philosophical sense.[6] What I dissent from in Sellars is his conviction that Kant's failure to discuss a formedness of "sense as such," independent of any involvement on the part of the understanding, is a ground for complaint.

Sellars's complaint is that "the characteristics of the representations of receptivity as such, which is what should *properly* be meant by the forms of sensibility, are never adequately discussed, and the so-called forms of sensibility become ever more clearly, as the argument of the *Critique* proceeds, forms of conceptual representations."[7] 'Never adequately discussed' seems an understatement; so far as I can see, Kant never so much as mentions what Sellars thinks he should have meant by the forms of sensibility. It is perhaps implicit in the "Aesthetic" that Kant thinks of sensation as the matter of empirical intuition (A20/B34), and thereafter he occasionally speaks of sensation as the matter of perception or of empirical knowledge (for example, A42/B59-60, A167/B209). But he never suggests that this matter has its *own* form as the matter it is, independently of its being formed into intuitions, perceptions, and empirical knowledge in the understanding-involving way that, as Sellars says, becomes ever more clearly Kant's concern as the *Critique* unfolds.[8]

---

[5] See *Science and Metaphysics*, p. 29. The Appendix on inner sense (pp. 230-38) offers a parallel move with respect to time as the form of inner sense.

[6] I have to be dogmatic here. On the philosophical question, I simply follow Sellars. On the question of fit with Kant (on which I am also following Sellars, though disagreeing with him over whether what he finds in Kant is a ground for complaint), see the footnote at B160. Kant there says that in the "Aesthetic" he represented the formal intuition, space, "as belonging merely to sensibility," but that was misleading; it does indeed "precede any concept," but it presupposes an operation of the understanding. I suppose a defender of the reading I am following Sellars in setting aside might claim that the form of outer intuition, which Kant here distinguishes from the formal intuition, could still, for all that the footnote says, be a topic for an autonomous inquiry into sensibility considered in abstraction from the understanding. But it is hard to see how spatiality, as the form of outer intuition, could be separated from the possibility of the "formal intuition," space itself as an object of intuition. If it cannot, the footnote implicitly instructs us not to suppose that the thesis that space is the form of outer sense is meant to be intelligible independently of the "Analytic."

[7] *Science and Metaphysics*, p. 30.

Sellars is convinced that a properly Kantian position requires forms of sense as such, forms of "sheer receptivity." Correctly in my view, he takes it that the "Aesthetic" does not consider such a topic. So something that should, he thinks, be fundamental to Kant's position is absent from the appropriate place in Kant's own presentation of it. (This conviction is reflected also in the other peculiarities Sellars finds in Kant's thinking: the counterpart to the mongrel conflation, and the "puzzling" abstractionist view of concept formation.) One is bound to wonder whether Kant can have so egregiously missed what is required for his own thinking. Perhaps a properly Kantian conception of outer sense needs no form of sense as "sheer receptivity," but only space in the role that Sellars, rightly in my view, takes Kant to attribute to it: as the form of outer intuitability on the interpretation according to which intuition involves the understanding as well as sensibility. That is, perhaps a below-the-line conception of sensibility need not have the transcendental role that Sellars credits to it. That is what I am going to urge.

III

Sellars quotes from a passage in the section of the *Critique* headed "The Clue to the Discovery of All Pure Concepts of the Understanding" (the so-called "Metaphysical Deduction"), where Kant says: "The same function which gives unity to the various representations *in a judgment* also gives unity to the mere synthesis of various representations *in an intuition*" (A79/B104-05).[9] But as I have noted, Sellars reads Kant as holding that when we take an intuition to represent its object as a *this-such*, what goes in place of 'such', at least in the case of intuitions at a certain basic level, need not yet be an expression of the corresponding concept as it might figure predicatively in a judgment. This means, I think, that Sellars cannot give the remark from the "Clue" its full weight.

Here I need to hark back to something I said in Lecture I, when I was trying to give the flavor of the Kantian idea that conceptual capacities have their paradigmatic actualization in judgment. If one judges, say, that there is a red cube in front of one, one makes a joint exercise of a multiplicity of conceptual capacities, including at least a capacity that would also be exercised in judging that there is a red pyramid in front of one and a capacity that would also be exercised in judging that there is a blue cube in front of one. This joint exercise of these capacities is not simply their being exercised in a single act of judgment. That would be equally true of a judgment that there is a red pyramid and a blue cube in front of one. The capacities have to be exercised with the right togetherness. If the judgment is to be that there is a red cube in front of one, the two capacities I have singled out have to be exercised with a togetherness that is a counterpart to the "logical" togetherness of 'red' and 'cube' in the linguistic expression of the judgment, 'There is a red cube in front of me'.

---

[8] Compare how one might think of bronze, say, as having its own form, independently of the forming of bits of it into statues or spearheads.

[9] For Sellars's citation, see *Science and Metaphysics*, p. 4.

We can connect this with the remark from the "Clue." This analogical specification of the mode of togetherness with which the two capacities I have singled out have to be exercised, if one is to be judging that there is a red cube in front of one, is a partial specification of the function that gives unity to the various representations in a judgment with that content, to put things in Kant's way.

In Lecture I, I used this conception of judgment as the basis for a parallel conception of the way in which perceptual (and specifically visual) experiences "contain" claims, as Sellars puts it in "Empiricism and the Philosophy of Mind." An ostensible seeing that there is a red cube in front of one would be an actualization of the *same* conceptual capacities that would be exercised in judging that there is a red cube in front of one, with the *same* togetherness. This captures the fact that such an ostensible seeing would "contain" a claim whose content would be the same as that of the corresponding judgment.

As actualizations of conceptual capacities with the appropriate togetherness, the judgment and the ostensible seeing would be alike. They would differ only in the way in which the relevant conceptual capacities are actualized. In the judgment, there would be a free responsible exercise of the conceptual capacities; in the ostensible seeing, they would be involuntarily drawn into operation under ostensible necessitation from an ostensibly seen object.

But since the two kinds of conceptual episodes are alike in respect of being actualizations of the appropriate conceptual capacities with the appropriate togetherness, the "logical" point of the remark from the "Clue" applies to ostensible seeings just as it applies to judgments. Following Sellars's lead, I have exploited an analogy between judging and claiming, in order to offer a partial specification of the function that gives unity to the various representations in a judgment that there is a red cube in front of one: a conceptual capacity corresponding to 'red' and a conceptual capacity corresponding to 'cube' have to be exercised with a togetherness corresponding to the togetherness of 'red' and 'cube' in 'There is a red cube in front of me'. Now the same specification is equally and by the same token a partial specification of the function that gives unity to the various representations in an ostensible seeing that there is a red cube in front of one. We can recast the remark from the "Clue" to say: the function that gives unity to the various representations in an ostensible seeing is the same as the function that gives unity to the mere synthesis of various representations in an intuition.

Any ostensible seeing will have more specificity to its content than just that there is a red cube in front of one, even if its content includes that.[10] For my purposes here, I can ignore much of what makes this true. But it matters for my purposes that in an ostensible seeing whose content can be partly specified as that there is a red cube in front of one, the apparent red cube will be *placed* more determinately than just somewhere or other in front of one. From the standpoint of the subject of such an ostensible seeing, its content will be expressible by saying something like 'There is a red cube there'. Here we have to imagine a use of 'there' that has a determinate significance by virtue of the subject's directing it in a specific way at the ostensible layout of the ostensibly seen environment. Similarly for a counterpart to such a use of 'there' in a nonovert conceptual occurrence that is to be understood on the model of making a claim by uttering those words (such as an ostensible seeing would be in the picture Sellars gives in "Empiricism and the Philosophy of Mind").[11]

Imagine, then, an ostensible seeing whose content is (in part) that there is a red cube *there*. (To imagine an ostensible seeing in these terms, we have to imagine it from the subject's point of view.) Now suppose this ostensible seeing is not a merely ostensible seeing, but a seeing. In that case, there *is* a red cube at the position the subject can mean by this kind of use of 'there' in an overt expression of the content of the experience in question, or by its counterpart in the nonovert conceptual occurrence that the experience is. In the conceptual occurrence that the experience is, the red cube that there actually is, given that the experience is a seeing, is itself directly in the subject's view. It is in the subject's view as *that red cube*. We can put it like that if we imagine ourselves into the subject's point of view; we have to imagine this use of 'that' as having a determinate significance by virtue of the same directedness at the (ostensible) layout of the (ostensibly) seen environment that we imagined as giving 'there' a determinate significance, when we imagined 'there' being used in specifying the content of the occurrence considered as an ostensible seeing that....

---

[10] See "Empiricism and the Philosophy of Mind," §22, p. 279 (151).

[11] Compare the notion of *conversio ad phantasmata* that Geach borrows from Aquinas in *Mental Acts*, pp. 65, 72, 74.

What I have arrived at here is a conception of a kind of representation (or at any rate *Vorstellung*) of an object that fits a standard Kantian characterization of intuitions: immediate sensible representations of objects (see, for example, A19/B33). The conception coheres with Sellars's insistence that intuitions in Kant's dominant sense belong above the line. 'Immediate' in a characterization of intuitions on these lines does not mean "not involving the understanding"; the intuitions that this characterization fits are not intuitions in the sense Sellars thinks Kant also needs, operations of "sheer receptivity." Sellars offers a different and better gloss on 'immediate' by urging that intuitions in the dominant Kantian sense are representations of *thises* (or *thats*); more fully, of *this-suches* (or *that-suches*), which makes it unavoidably clear that even though they are immediately of objects, such representations already involve the understanding.[12] The remark from the "Clue" points to how we can conceive intuitions in this sense as actualizations of conceptual capacities with a suitable "logical" togetherness.

An ostensible seeing is an actualization of conceptual capacities with a specific "logical" togetherness. What makes it an ostensible seeing, as opposed to a conceptual episode of some other kind (for instance, a judgment), is that this actualization of conceptual capacities is a conceptual shaping of sensory (and in particular visual) consciousness. The remark from the "Clue" says that an intuition is characterized by the same "logical" togetherness. If an ostensible seeing is a seeing, then the conceptual shaping of visual consciousness that constitutes it, those very conceptual capacities actualized in visual consciousness with that very "logical" togetherness, constitute—looked at, as it were, from a different angle—an intuition: an immediate presentness of an object to sense. A seeing that... is a seeing *of an object*, at least if its content is of the sort that figures in the example I have been working with. To apply what Kant says in the "Clue" to my example: the function that gives unity to the various representations in a judgment whose content we can imagine capturing from the subject's viewpoint as that there is a red cube *there* (the function that unites the various conceptual capacities exercised in such a judgment), or (this comes to the same thing) the function that gives unity to the various representations in an ostensible seeing with that same content (the function that unites the various conceptual capacities actualized in such an ostensible seeing), is the same function that—in the sort of case in which there is an intuition; that is, in the sort of case in which the ostensible seeing is a seeing—gives unity to the mere synthesis of various representations in an intuition of *the red cube there* or *that red cube*, to speak again from an imagined occupation of the subject's viewpoint.[13]

---

[12] Compare *Science and Metaphysics*, p. 3.

Here the fact that, say, 'cube' figures in a specification of the content of an intuition—the intuition represents its object as that red *cube*—reflects the fact that for one to be the subject of such an intuition is in part for there to be actualized in one's sensory consciousness the very same *conceptual* capacity—possession of the concept of a cube—whose exercise would partly determine the predicative element in the content of a judgment whose content we could specify, with that imagined occupancy of the subject's viewpoint, in the form 'That is a red cube'. In fact, the actualization of the relevant conceptual capacity in the intuition *is* an actualization of it in a conceptual occurrence whose content is, so to speak, judgment shaped, namely, a seeing (a seeing that...) whose content is that there is a red cube *there*—as we can put it with the same imagined occupancy of the subject's viewpoint. This seeing that... , in describing which we explicitly place an expression for the concept in question in a predicative position, is the very same conceptual occurrence—an actualization of the same conceptual capacities with the same "logical" togetherness—as the intuition.

So when we say that an intuition represents its object as a cube, the word 'cube' does not signal a merely protoconceptual contribution of the understanding to the constitution of the intuition, as it can in Sellars's Kant. Sellars insists, rightly in my view, that intuitions in the dominant Kantian sense already involve the understanding. I have been expressing this by speaking of intuitions not as *conceptual* shapings of sensory consciousness but as shapings of sensory consciousness *by the understanding*. That was to make room for the fact that in Sellars's reading, although the understanding, the faculty of concepts, is operative in the constitution of intuitions, concepts proper are not, at least in the case of intuitions at a certain basic level. But with the different reading I have arrived at, I can drop the circumlocution, and speak of intuitions, just like seeings that..., simply as *conceptual* shapings of sensory consciousness. Visual intuitions *of objects* simply are seeings that..., looked at as it were from a different angle. There is no opening here into the abstractionist picture of the formation of basic empirical concepts that Sellars finds in Kant.

---

[13] How can something describable as a seeing that there is a red cube *there*, an intuition of *that* red cube, be a conceptual episode, given the characteristic Kantian association of concepts with generality (see, for example, A320/B377)? There is no problem in the idea of a general conceptual capacity, an actualization of which is indicated by the fact that a particular experience can be described in those terms (where the description indeed exploits the particularity of a subject's experiential situation). This capacity (a capacity to mean determinate places by utterances of 'there' or nonovert counterparts, and to mean determinate objects by utterances of 'that ' or nonovert counterparts) is not restricted to the particular actualization of it we are imagining

In the passage in the "Clue," Kant speaks of "the mere synthesis of various representations" in an intuition. The insertion of 'mere synthesis' implicitly differentiates the unity of an intuition from the unity of a judgment. It is plausible that this is connected with a passage Sellars also quotes, from just before the remark I have been considering, in which Kant speaks of synthesis as "the mere result of the power of imagination, a blind but indispensable function of the soul" (A78/B103). Sellars might cite this in support of his view that Kant envisages intuitions that are only protoconceptual, even though they already involve the synthetic powers of the understanding, so that they can be a source from which concepts proper can be derived. But these remarks of Kant's are perfectly intelligible on the different reading I am giving. The point is simply that it does not take cognitive work for objects to come into view for us. Mere synthesis just happens; it is not our doing, unlike making judgments, deciding what to think about something. This is quite consistent with holding that objects come into view for us in actualizations of capacities that are fully conceptual, capacities whose paradigmatic mode of actualization is in the exercise of cognitive responsibility that judging is.

IV

I am following Sellars in taking it that we can express the content of a Kantian intuition by a phrase such as 'that red cube'. We might suppose that conceptual occurrences whose content can be given like that, with a phrase that is less than a whole sentence, are essentially potential ingredients in some more extensive conceptual goings-on—say, in the judgment that that red cube is too big to fit in the box. (As before, in order to make sense of this use of 'that', we have to imagine ourselves into the viewpoint of the subject of a specific case of the kind of conceptual occurrences in question.)

I do not want to dispute this suggestion. But the point that matters for my purposes is that these conceptual occurrences, whose content we can express with mere phrases, can also be conceived in a way that equips them—the very same conceptual occurrences—with judgment-shaped contents: specifically, for my example, the judgeable content that there is a red cube *there*. (Again, we have to imagine ourselves into the subject's viewpoint in order to deal with this use of 'there'.) The other thought—that intuitional content is essentially a fragment of judgmental content—would imply that the ability to have objects come into view for one is essentially dependent on the ability to make judgments, and that is indeed an implication of the position I am finding in Kant. But the point I am stressing yields not just that but also something more radical: that an actualization of the capacity to have objects come into one's view is itself already an actualization of the capacity to have occur in one's life occurrences with the sort of content that judgments have, not just an element in such an actualization.

So far as that goes, we might suppose the capacity to be a subject of actualizations of conceptual capacities with judgmental content just happens to be, but might not have been, sometimes actualized in the shape of intuitions: occurrences in which objects come into view, literally in the case of visual intuitions. But this cannot be Kant's view; that would fly in the face of his insistence that intuitions are indispensable if thought is to be contentful at all (for example, A50-52/B74-76). Kant's view must rather be something like this: the very idea of a conceptual repertoire is the idea of a system of capacities that allows, as it were at the ground level, for actualizations in which objects are immediately present to the subject.

This suggests a new twist to the nontraditional empiricism that, in Lecture I, I found in Sellars's discussion of the metaphor of foundations of knowledge, in "Empiricism and the Philosophy of Mind." Sellars's picture there is already deeply Kantian, and we can now see how to make it even more so.

As I pointed out, Sellars does not dispute the empiricist thought that everything else in a world view depends on perceptual knowledge, in a logical dimension in which we are moving when we relate beliefs to their credentials. He transforms the empiricist picture by adding another logical dimension in which a different kind of dependency is traceable in the opposite direction. What he insists is that the very idea of perceptual knowledge, and more generally the very idea that perceptual experiences, whether knowledge yielding or not, "contain" claims, so that they can be so much as putatively knowledge yielding, depend in this other logical dimension on the fact that the claims "contained" in perceptual experiences have their places in a world view. This dependence of foundations (now at best awkwardly so called) on superstructure is a transcendental matter. The claim is that we can intelligibly credit perceptual experiences with objective purport only in virtue of how the conceptual apparatus that constitutes their objective purport fits into the world view that is, in the other logical dimension, grounded on the deliverances of experience. But the downward dependence, the dependence of superstructure on foundations, is still narrowly epistemological.

The new twist is that, with the conception of Kantian intuitions I am urging, we can put into the picture a downward dependence that is not narrowly epistemological but, like the upward dependence that is already in Sellars's picture, transcendental, a matter of requirements for it to be intelligible that the picture depicts directedness at objective reality at all. Kant implicitly says that thought without intuitions would be empty.[14] We can now see that his point is not—at least not in the first instance—to insist that concepts must be capable of figuring in, say, judgments that can be *grounded* in experiences (which, to play this role, would have to be conceived as possessing judgeable content, for instance as seeings that...). The transcendental requirement is that it must be intelligible that conceptual activity has a *subject matter*. And Kant's thought is that this is intelligible only because we can see how the very idea of a conceptual repertoire provides for conceptual states or episodes in which a subject matter for conceptual activity is sensibly present, plainly in view in actualizations of capacities that belong to the repertoire.[15]

Not, of course, that we cannot direct thought at objects that we are unable to bring into view, perhaps because they are too small or too far away. But thought so directed is carried to its object, so to speak, by theory. The ultimate credentials of theory must lie in experience. And we can make sense of the idea, which is so far epistemological in the narrow sense, that the ultimate credentials for theory lie in experience—we can make sense of experience as made up of, for instance, seeings that... —only because we can make sense of experience as bringing objects into view. Concepts, which make thought what it is, can intelligibly be what they are—thought can intelligibly be of the objective at all—only because we can see how there can be conceptual occurrences in which objects are manifestly there for thinkers, immediately present to their conceptually shaped sensory consciousness. But, equally, there can intelligibly be such conceptual occurrences only because we can see how thought can also be related to its subject matter in a way that is mediated by theory; this is Sellars's upward dependence applied to the relation between intuitions and world view.

---

[14] This is implicit in the remark: "Thoughts without content are empty, intuitions without concepts are blind" (A51/B76).

[15] This corrects the picture I gave, meaning it to be Kantian, in *Mind and World*. There I took it that 'object', in the Kantian idea that intuitions are of objects, just meant 'objective somewhat' (including, for instance, states of affairs) I now think it means something much closer to what 'object' means in the standard translations of Frege.

That last remark opens into a topic I have so far passed over in silence. Kant's "Clue" is only a clue, to the discovery of the so-called pure concepts of the understanding; when he makes the remark I have been exploiting, we are still waiting for the transcendental deduction of the "pure concepts." So, even supposing I am on the right track in the conception of intuitions I have elicited from the "Clue," Kant's thought must be that work still needs to be done for us to be entitled to the idea of an immediate presentness of objects to subjects in intuition. We still need to understand how the categories make experience possible.

This is a large and complex matter, which I cannot go into in these lectures. (So far as I can see, the categories are in a similar way not explicitly present in the parallel region of Sellars's partial reading of Kant, which is what I am aiming to exploit and improve on.[16]) What I have already said, however, yields a hint as to the direction in which this topic lies from the position we have reached. I have found a conception of intuitions suggested in the "Clue," and I have been importing that conception of intuitions into a variant of the picture Sellars gives when he reflects on the metaphor of foundations. In doing that, I have been tacitly pointing to a place for something on the lines of the categories. In order to be entitled to see conceptual activity as having objective purport, we have to see how actualizations of conceptual capacities include intuitions. But—the new version of Sellars's upward dependence—we can make sense of objects coming into view in intuitions only because we can see how objects fit into a view of the world. Something like the categories, and the principles Kant connects with them, would figure in giving substance to that thought.[17]

---

[16] But see his paper "Some Remarks on Kant's Theory of Experience," in *Essays in Philosophy and Its History* (Dordrecht: Reidel, 1974).

V

Sellars thinks the transcendental role of sensibility, in properly Kantian thinking, is to supply manifolds of "sheer receptivity" to guide conceptual representations. At one point, he suggests that Kant needs this picture if he is to "avoid the dialectic which leads from Hegel's *Phenomenology* to nineteenth-century idealism."[18] This remark is instructive.

Sellars is invoking Hegel as a bogeyman: as someone who, by failing to acknowledge any external constraint on thought, makes it unintelligible how what he is picturing can be directed at what is independently real, as it must be if it is to be recognizably thought at all. But this takes no account of the fact that Hegel thinks he finds his notion of Reason, moving freely in its own sphere, adumbrated precisely in Kant's attempts to characterize the interpenetration of sensibility and understanding, in the first *Critique* and the *Critique of Judgment*.[19] Hegelian Reason does not need to be constrained from outside, precisely because it includes as a moment within itself the receptivity that Kant attributes to sensibility.

Now, perhaps the conception of Kantian intuitions that I have arrived at is a way to begin bringing this difficult conception down to earth. Sellars's idea is that for thought to be intelligibly of objective reality, the conceptual representations involved in perceptual experience must be guided from without. Indeed they are, I can say. But there is no need for manifolds of "sheer receptivity" to play this guiding role. In a way we are now equipped to understand, given the conception of intuitions adumbrated in the passage from the "Clue," the guidance is supplied by *objects* themselves, the subject matter of those conceptual representations, becoming immediately present to the sensory consciousness of the subjects of these conceptual goings-on.

---

[17] Sellars says "there is an important sense in which one has *no* concept pertaining to the observable properties of physical objects in Space and Time unless one has them all—and, indeed, as we shall see, a great deal more besides" ("Empiricism and the Philosophy of Mind," §19, p. 275 (148)). 'As we shall see' points forward to the discussion of the metaphor of foundations, where the claim becomes that one has no concept of the observable unless one knows a great deal about the world. We can put Sellars's thought here in Kantian terms: acquiring one's first conceptual capacities is necessarily acquiring many conceptual capacities, interlinked in such a way that the totality amounts to a conceptual repertoire that exemplifies the necessary forms of the understanding. It comes to the same thing to say that acquiring one's first conceptual capacities is necessarily acquiring a world view that conforms to the associated principles of pure understanding. (Of course, this allows divergence from Kant over what these forms and principles are.)

[18] *Science and Metaphysics*, p. 16. See also p. 29, where, in connection with Kant's failure to distinguish what he supposedly needs the forms of sensibility to be from "the 'forms' of that which is represented by the intuitive conceptual representations which are 'guided' by receptivity," Sellars says that "no sooner had [Kant] left the scene than these particular waters were muddied by Hegel and the Mills, and philosophy had to begin the slow climb 'back to Kant' which is still under way." Presumably, the idea is that Hegel tried to do everything in terms of what is above the line, without the transcendentally needed guidance by "sheer receptivity," whereas the Mills reverted to an empiricistic version of the Myth of the Given, trying to do everything ultimately in terms of what is below the line.

[19] I have learned here from Béatrice Longuenesse, "Point of View of Man or Knowledge of God: Kant and Hegel on Concept, Judgment and Reason" (delivered at the University of Pittsburgh, January 1997).

Sellars's own imagery for expressing his sense of the need for external constraint—his talk of guidance and the like—actually fits this constraint by subject matter better than it fits Sellars's candidate, constraint by "sheer receptivity."

Any faithful student of "Empiricism and the Philosophy of Mind" must be made uneasy by finding Sellars, in *Science and Metaphysics*, saying that states or episodes below his line *guide* states or episodes above it. This seems dangerously close to a lapse into the Myth of the Given, by Sellars of all people.[20]

At one point, speaking of the operations of the understanding in perception, in the guise of the productive imagination, Sellars is led to say, even more poignantly, that the transcendentally posited manifold of nonconceptual sensory impressions "is an independent factor which has a strong voice in the outcome" (*op. cit.*, p. 16). This runs the same risk of seeming to lapse back into the idea of the given. But even apart from that, this image of voice is difficult to cash out. When the manifold of "sheer receptivity" is playing the role Sellars tries to capture here with the image of voice, it is not speaking to *us*; that is a natural metaphor for Sellars's own thought, which I discussed in Lecture I, that when it is playing that role it is not apperceived. When a manifold of "sheer receptivity" does speak to us, on this natural interpretation of the image, it is no longer playing its

[20] Starting here, one can see how Brandom's conception of observational claims and knowledge, in chapter 4 of *Making It Explicit*, might be represented, not as radically non-Sellarsian, but as a charitable reading of Sellars's basic intentions. On this view, Sellars's own wish to keep sensory consciousness in the picture is a vestige of an archaic and risky philosophical outlook, which he himself undermines in his attack on the Myth of the Given. A hygienic replacement is the bare idea of reliable differential responsive dispositions. On this view, sentience is a mere detail of the causal connection between the responses and what they respond to; taking it to have more philosophical importance than that merely courts pre-Sellarsian dangers, as Sellars himself surprisingly does in *Science and Metaphysics*. I agree with Brandom that Sellars's thinking needs a charitable reading; but I am offering a way to preserve Sellars's Kantian thought that sensibility is transcendentally important, without running the risks that Brandom rightly wants to avoid.

transcendental role. To whom or what, then, is it supposed to speak when, in playing its transcendental role, and hence not being apperceived, it is having its say in the outcome of the operations of the understanding in perceptual experience? Perhaps to the understanding at its work? But if the manifold of "sheer receptivity" speaks to our understanding but not to us, we have an awkward separation of our understanding from ourselves, as if the understanding were a distinct cognitive subject within a person, doing its work as it were behind our backs.

But suppose we take it that the external constraint Sellars sees to be required is exerted, in intuition, by objects themselves, the subject matter of the conceptual representations involved in perception. Now the image of voice fits more easily. A seen object as it were invites one to take it to be as it visibly is. It speaks to one; if it speaks to one's understanding, that is just what its speaking to one comes to. "See me as I am," it (so to speak) says to one; "namely, as characterized by *these* properties"—and it displays them.

Of course this comparison of images cannot settle the question raised by Sellars's invocation of Hegel. After all, in "Empiricism and the Philosophy of Mind" Sellars himself exploits an image in which conceptual representations in perception are "evoked" or "wrung" from us *by objects*. His transcendental thought is that we can be entitled to this image of external constraint—a version of the idea Kant expresses by saying objects are what prevent our cognitions "from being haphazard or arbitrary" (A104)—only if we acknowledge that the conceptual representations we want to think of in those terms are guided by "sheer receptivity."

Sellars's thinking here is bound up with a doctrine of his about the relation between "the scientific image" and "the manifest image," which he puts in "Empiricism and the Philosophy of Mind" by saying: "*speaking as a philosopher*, I am quite prepared to say that the commonsense world of physical objects in Space and Time is unreal—that is, that there are no such things."[21] According to this doctrine, the red cubes and so forth that are, apparently, immediately present to us in intuitions do not really exist. So they cannot be what guide our conceptual representations in perception from outside. Presumably we must suppose that our conceptual representations are guided by the items that the scientific image substitutes for these merely apparent objects—swarms of colorless particles or whatever. But this real guidance cannot have the immediacy of the guidance by red cubes and the like that figures in the manifest image. So if that apparent immediate guidance by objects is to figure in the transcendental project of showing how it is that our conceptual activity is intelligibly conceptual activity at all, directed, as that requires, toward objective reality, then we have to reconceive it—transcendentally, or speaking as philosophers—as mediated guidance by genuinely real, nonsensible objects. If we want to conceive this mediated guidance by real objects as immediate guidance by something real, it can only be as guidance by the sensory goings-on out of which the productive imagination constructs the red cubes and the like that figure in the manifest image as the immediate objects of intuition.

---

[21] §42, p. 303 (173). See "Philosophy and the Scientific Image of Man," in *Science, Perception and Reality* I am indebted to Anders Weinstein for insisting on the relevance of this aspect of Sellars's thinking.

Sellars is committed to this as a reading of Kant's distinction between appearances, as immediate objects of intuition, and things in themselves.[22] Here I think his often wonderful attunement with the spirit of Kant's thinking deserts him. This Sellarsian picture cannot accommodate Kant's insistence that the things in themselves that matter for his thinking about empirical knowledge are the very same things that make their appearance in intuition.[23] I think it comes to the same thing to say: Sellars's idea that the red cubes and so forth of common sense do not really exist is philosophically misguided. I cannot properly justify these beliefs in these lectures, but I hope to have begun on making it plausible that there is an alternative to what Sellars represents as compulsory.

---

[22] See *Science and Metaphysics*, chapter 2.

[23] See, for example, Bxxvii, where Kant speaks of "the distinction, which our Critique has shown to be necessary, between things as objects of experience and those same things as things in themselves." When we speak as philosophers, we do not start to speak of a new range of objects, genuinely real as the objects of the manifest image were not. We speak of the same objects, under a special mode of consideration in which we abstract from the way in which the objects figure in our world view. Sellars reads Kant as a scientific realist manqué; in Sellars's view, had Kant only been sophisticated about the possibilities for scientific concept formation, he would have cast the objects of the scientific image, which are distinct from the objects of the manifest image, in the role of things in themselves. But for Kant objects as they appear in the scientific image would be just another case of objects as they appear, with a transcendental background for that conception just as necessary here as anywhere. Sellars's attempt to be responsive to Kantian transcendental concerns goes astray in his idea that an appeal to science could do the transcendental job; here Sellars's scientism is seriously damaging.

I am here correcting the two-worlds picture of Kant that I presupposed in *Mind and World*. But note that what Kant insists on in Bxxvii is an identity of things as they appear in our knowledge and "those same things as things in themselves"; not "those same things as they are in themselves." Wording on the lines of the latter pervades, for example, Henry E. Allison's non-two-worlds reading, in *Kant's Transcendental Idealism* (New Haven: Yale, 1983). Considering things as things in themselves is considering the very things that figure in our knowledge, but in abstraction from how they figure in our knowledge. That is not to say, considering them as possessing, unknowably to us, other properties than those they appear as possessing in our knowledge of them. It is not to distinguish *how they are* in themselves from how they appear in our world view. With this latter construal of things in themselves, it seems to me that the non-two-worlds reading might just as well be a two-worlds reading. The picture still involves two realms of fact, one knowable by us and one unknowable by us; it does not undermine the damage this does to say that the same objects figure in both.

It might be tempting to defend Sellars on these lines. If we do not acknowledge a transcendental need for guidance by "sheer receptivity," then we face a dilemma. On one horn, our attempt to make sense of conceptual activity as having objective purport degenerates into an "idealistic" fraud; the so-called reality toward which we see our so-called conceptual activity as directed is a mere projection of the activity. On the other horn, we fall into an impossible transcendental realism, which we can make vivid by coalescing the image of objects speaking to us with an image of Richard Rorty's; on this horn, we picture objects as speaking to us in the world's own language.[24]

But this dilemma does not threaten the position I am urging.

I do not picture objects as speaking to us in the world's own language. Objects speak to us, in the metaphor that fits the position I am urging, only because we have learned a human language. We can play with the image of objects speaking to us in a language we know (say English), as I did a moment ago. But, less fancifully put, the point is that objects come into view for us only in actualizations of conceptual capacities that are ours. To entitle ourselves to this, we must acknowledge whatever we need to acknowledge for the conceptual capacities to be intelligibly ours. The fantasy of conceptual capacities that belong to the world itself is not to the point.

That does not land me on the other horn of the dilemma, according to which the so-called objects can only be projections of our thinking. Objects come into view for us in actualizations of conceptual capacities in *sensory* consciousness, and Kant perfectly naturally connects sensibility with receptivity. If we hold firm to that, we can see that the presence of conceptual capacities in the picture does not imply "idealism," in the sense in which Sellars means invoking "idealism" to frighten us. If we conceive subjects as receptive with respect to objects, then, whatever else we suppose to be true of such subjects, it cannot undermine our entitlement to the thought that the objects stand over against them, independently there for them.

---

[24] *Op. cit.*, p. 298: "successfully representing according to Nature's own conventions of representation." For a Sellarsian formulation close to this imagery, see "Empiricism and the Philosophy of Mind," §34, p. 296 (167).

LECTURE III: INTENTIONALITY AS A RELATION

Sellars shows us how to understand visual experiences as ostensible seeings, occurrences in a subject's visual life that "contain" claims about an ostensibly visible region of objective reality. That they "contain" claims is the same fact as that they are conceptual occurrences, actualizations of conceptual capacities with a suitable "logical" togetherness. In that respect, they are like judgments. But they are unlike judgments in the way in which they "contain" their claims. Judgments are free exercises of conceptual capacities with a suitable togetherness. But in an ostensible seeing whose content includes that of a given judgment, the same conceptual capacities are actualized, with the same togetherness, in a way that is ostensibly necessitated by the objective reality that is ostensibly seen. A visual experience is a case of being under the *visual impression* that things are thus-and-so in the ostensibly visible environment.

This picture of visual experiences as conceptual shapings of visual consciousness is already deeply Kantian, in the way it appeals to sensibility and understanding so as to make sense of how experiences have objective purport. But Sellars thinks that to be fully Kantian, the picture needs a further element. He thinks this idea of conceptual shapings of visual consciousness is something we can entitle ourselves to only by means of a transcendental postulation, according to which these conceptual goings-on are guided by manifolds of "sheer receptivity": occurrences in visual consciousness that are not conceptually shaped. Only so, Sellars thinks, can we legitimately take perception to yield conceptual representations of objective reality. If we do not acknowedge that the conceptual goings-on in perception are guided by "sheer receptivity," then no matter what we do toward purporting to equip ourselves with an object of perceptual awareness, the supposed object can be no better than a projection of our mental activity. This is what Sellars implies by invoking the idealism that Hegel made room for, as a pitfall awaiting Kant if he fails to credit sensibility with the role Sellars thinks it must have.

In holding that a properly Kantian view requires this guidance by "sheer receptivity," Sellars commits himself to taking a dim view of some features of the first *Critique*. The problems in Kant's own exposition, as Sellars sees them, come to a head with Kant's thesis that space is the form of outer sense. What space informs is outwardly intuitable phenomena, and it becomes increasingly clear as we progress through the *Critique* that outer intuition involves the understanding as well as sensibility. Focusing on space as informing outer intuitability, Kant fails to discuss what according to Sellars ought to be a crucial topic for him, the form of the manifolds of "sheer receptivity" that Sellars thinks must guide conceptual representations of outer reality in perception. He says Kant does not "adequately" discuss this supposedly crucial topic, but really he does not discuss it at all.

Now obviously this mismatch between what Kant wrote about forms of sense and what Sellars thinks he ought to have written should give us pause. By itself, however, this carries little weight. No doubt the mismatch gives Sellars pause too; but he self-consciously claims to see better than Kant what Kant ought to have written. His thought is that there must be guidance by "sheer receptivity" if we are to be entitled to the basically Kantian idea that, by conceiving sensibility as shaped by the understanding, we make it intelligible how sensory consciousness can be directed toward objective reality, and thereby how thought in general can have objective purport. If Sellars is right about that, then he is right that, in failing to discuss forms of "sheer receptivity," Kant overlooks something crucially required by fundamental features of his own thinking. So the real question is whether Sellars is right about the necessity for guidance by "sheer receptivity."

In Lecture II, I began urging that Sellars is wrong about that. Kant conceives intuitions as representations in which objects are immediately present to subjects. From a remark in the section called "The Clue to the Discovery of All Pure Concepts of the Understanding," I elicited this suggestion: intuitions are conceptual occurrences that exemplify exactly the "logical" togetherness, on the part of actualizations of conceptual capacities in sensory consciousness, in terms of which we can make it intelligible that ostensible seeings "contain" claims about the objective environment. In fact, visual intuitions just are the actualizations of conceptual capacities, with the requisite togetherness, that constitute those ostensible seeings that are seeings. If an ostensible seeing that... is a seeing that... , it is itself an intuition, at least if the content of the ostensible seeing deals with an ostensible object—as in the case of an ostensible seeing that there is a red cube at a position in the ostensibly visible environment that one can single out as *there*.

Sellars thinks the conceptual representations in perception must be guided by manifolds of "sheer receptivity," because he thinks that only so can we make it intelligible to ourselves that conceptual occurrences in perceptual experience—and thereby ultimately thought, conceptual activity, in general—are constrained by something external to conceptual activity. As he sees, we need such external constraint in our picture if we are to be entitled to take it that conceptual activity is directed toward an independent reality, as it must be if it is to be intelligible as conceptual activity at all. But I suggested that once we understand how objects can be immediately present to conceptually shaped sensory consciousness in intuition, we can take this need for external constraint to be met by perceived objects themselves. The transcendental task is entitling ourselves to see conceptual activity as directed toward a reality that is not a mere reflection of it. To discharge that task, we need not see conceptual representations in perception as externally constrained by anything except the relevant elements of the very independent reality toward which we are in the course of entitling ourselves to see conceptual activity, in general, as directed. There is a kind of circularity here, but not one that should make it look as if the putatively constraining objects can only be projections of what we are trying to see as conceptual activity—in vain, if we could not do better than this. The actualizations of conceptual capacities that we are focusing on when we do this transcendental work are shapings of *sensory* consciousness, and thus of what Kant describes, with an obvious appropriateness, in terms of receptivity. That ensures that the objects we are entitling ourselves to see as present to subjects in intuition are genuinely independent of the subjects.

In "Empiricism and the Philosophy of Mind," Sellars himself talks of conceptual representations evoked by perceived objects. But he would relegate the appropriateness of such talk to the manifest image, and he would urge that we need to give it a transcendental vindication by showing how it correlates with the scientific image. Otherwise the putative objective purport that figures in the manifest image would be a mere illusion; the apparently perceived objects—such things as red cubes—do not really exist. Now of course there are familiar supposed grounds, of a scientistic sort, for denying independent reality to the immediate objects of perception, and attributing it only to their counterparts in the scientific image. I think these grounds are unconvincing, but I am not going to consider them in this lecture.[1] What I am going to discuss is a different and I think more interesting feature of Sellars's thinking: a different way to understand why, for him, guidance by the immediate objects of intuition cannot itself figure in transcendentally vindicating the very idea of objective purport.

---

[1] Finding them unconvincing does not require me to debunk the scientific image, but only to question its claim to contain all the real objects there are

I

In the original version of "Empiricism and the Philosophy of Mind," Sellars implies that seeings are the veridical members of the class of ostensible seeings (*op. cit.*, §7, pp. 260-01 (133-34); cf. §22, pp. 279-80 (151-52)). For the reprinting in *Science, Perception and Reality*, he adds a couple of footnotes suggesting that if an ostensible seeing is to be a seeing, it must not only be veridical, but in addition the subject must know that the viewing circumstances are normal.[2]

Now it was surely wrong to imply that veridicality is all it takes for an ostensible seeing to be a seeing. Consider a case in which someone is screened off from a red cube by a successful *trompe l'oeil* painting in which an indistinguishable red cube is depicted as being precisely where the unseen red cube actually is. Here we have a veridical ostensible seeing that is not a seeing. But Sellars's attempt to correct this mistake, in the added footnotes, seems unhappy. Surely one might have occasion to say: "I now realize I was seeing a red cube, although at the time—because I thought the circumstances were abnormal—I did not realize it." What is perfectly intelligibly claimed here is that the case was one of seeing, even though the subject did not know that the viewing circumstances were normal. What matters is that the circumstances should be normal, not that the subject should know they are.

Sellars's first thought would have been better put by saying that an ostensible seeing is essentially an ostensible *seeing*. His mistake was just to forget that being nonveridical is not the only way an ostensible seeing can be merely ostensible. Sellars's second thought poses a risk that is definitely avoided if we correct his first thought in that way; the invocation of normality encourages supposing we can build up to the notion of seeing by adding conditions to an independent notion of visual experience.

I extracted from Kant's "Clue" the idea that, if an ostensible seeing that... is a seeing that... , the very actualization of conceptual capacities that accounts for its "containing" its claim also constitutes—at least if the content of the claim deals with an ostensible object—its being an intuition, in which an object is immediately present to the subject. Now we debar ourselves from this notion of immediate presentness of objects to subjects if we let it seem that a seen object would have to figure in the content of a conceptual occurrence that is a seeing of it as, for instance, occupying a position at the outer end of a causal chain that generates the subject's current experiential situation in some suitably designated way. And Sellars's second thought suggests just that. It suggests that seeings that... would need to "contain" not just claims about the environment but also claims to the effect that the subject's experience is "normally" related to the ostensibly seen environment (this being part of what the subject is supposed to know in enjoying an experience of the relevant kind). That introduces a mediation that would threaten our ability to take these same conceptual occurrences to be intuitions, immediately of objects, as the remark from the "Clue" suggests we should be able to.

---

[2] §22, pp. 151, 152 (in *Science, Perception and Reality*). I am grateful to Paul Coppock for drawing my attention to these footnotes.

This connection of immediacy with an absence of extra conceptual content matches an element in Gareth Evans's account of perceptually demonstrative thoughts (*op. cit.*, chapter 6). Evans says demonstrative thoughts in the most basic sense are carried to their objects by an "information-link" that connects the objects to the subjects, rather than by a thought of the link. (Of course the counterpart I have arrived at is compatible with insisting, as Evans does, that thought can go directly to its object like this only against the background of a richly situating self-consciousness on the part of its subject.)

And we can take this correspondence with Evans further. The actualizations of conceptual capacities that constitute ostensible seeings can amount to intuitions, cases of having objects immediately present to one, only if the ostensible seeings are seeings. Of course merely ostensible seeings are ostensible *seeings*, so that—at least if their content deals with ostensible objects—they ostensibly constitute intuitions. But the mere appearance of an intuition is just that; it is not an actual intuition. I have been following Sellars in connecting the immediacy of intuitions with their being "representations of *thises*."[3] Now if I put the point that merely ostensible seeings afford mere appearances of intuitions in terms of this connection between intuitions and demonstratives, it amounts to this: Kant's conception of intuitions embodies a version of Evans's thesis that perceptual demonstrative content is *object dependent*. If one is under the illusion of being perceptually confronted by an object, then one is liable to a counterpart illusion that there is available to one, for employment in conceptual activity, content expressible by a perceptual demonstrative reference to the supposed object—the content one might think one could express, in such a situation, by using a phrase such as 'that red cube'. This is just what the immediacy of intuitions comes to; if there can be conceptual shapings of sensory consciousness in which objects are immediately present to subjects, then illusions that objects are present to one in that way, which obviously can happen, are at the same time illusions about the contents of one's conceptually shaped consciousness.

---

[3] *Science and Metaphysics*, p. 3.

This imputation of an illusion of content is often found counterintuitive (to use an indispensable term in spite of its awkwardness in this context). But the remark from the "Clue" makes it doubtful that this complaint has any substance. In a merely ostensible seeing that there is, say, a red cube at a position one can mean by a use of 'there', there are *actualized* in one's visual consciousness conceptual capacities corresponding to the presence of the words 'red', 'cube', and 'there' (in a use that exploits one's experiential situation) in a verbal expression of the experience's content. None of that conceptual content is an illusion. In the language of the remark from the "Clue," there is a function that does indeed give unity to the various representations in the content of the ostensible seeing, or rather this part of its content (since any actual ostensible seeing will have more to it). The content in question is the same as the content of a judgment the subject might express by saying 'There is a red cube *there*'. What is illusory is just the appearance that the same function also gives unity to a synthesis of the same representations in an intuition. The relevant function—the "logical" togetherness with which the relevant conceptual capacities are actualized—certainly *seems* to give unity to a synthesis of the representations in an intuition; that is to say that there *seems* to be a red cube immediately present to the subject. But since there is no such red cube—since the ostensible seeing is merely ostensible—this seeming intuitional unity is a mere semblance of an intuitional unity; that is to say that there *merely* seems to be a red cube immediately present to the subject.

II

Exploiting the way Sellars suggests we should understand the immediacy of intuitions, I have arrived at a notion of object-dependent conceptual content expressible in demonstrative reference. But the notion I have arrived at is radically alien to Sellars. This yields a way to bracket Sellars's doctrine that the objects of the manifest image do not really exist, but still have an explanation for his inability to countenance the transcendental role I have suggested we can attribute to this conception of intuitions.

The point is that the picture depicts intuitions as, qua conceptual occurrences of a certain kind, *related* to objects. The idea that there is only an illusion of content in certain situations simply makes vivid the relational character of the conceptual occurrences that are intuitions. The idea is that for a conceptual episode to possess intuitional content just is for it to stand in a certain relation to an object; so if there is no object suitably related to a conceptual episode, then there is no such relation, and accordingly no such content. (Of course there is still a conceptual episode, an ostensible perceiving.) But it is central to Sellars's thinking that elements in the conceptual order can stand in content-involving or semantical relations only to elements in the conceptual order, not to elements in the real order. He thinks this "non-relational character of 'meaning' and 'aboutness'" is "the key to a correct understanding of the place of mind in nature" (*ibid.*, p. ix).

III

How can meaning and aboutness be nonrelational? In expounding this Sellarsian thought, it is easiest to begin with the semantical character of elements in language, in the sense in which language is a repertoire for overt linguistic acts.

By saying what an expression means or stands for, we capture the expression's potential for making it the case that a linguistic act in which it occurs has a specific directedness toward extralinguistic reality.[4] In this context, the thesis we are concerned with is that such semantical statements do not relate the expressions they deal with to elements in extralinguistic reality.

On the face of it, the forms '...means—' or '...stands for—' are relational. If the expressions that figure on the right-hand sides of statements of such forms were used in the way they are used in ordinary (nonsemantical) discourse, we could take that appearance at face value; we could suppose that the expressions refer to certain elements in extralinguistic reality, those toward which ordinary uses of them enable linguistic acts to be directed, so that the statements could be taken to assert relations between these elements in extralinguistic reality and the expressions mentioned on their left-hand sides. But according to Sellars, the expressions on the right-hand sides of these statements are not used, at any rate not in that ordinary way. And they are not exactly, or not merely, mentioned either. If they were merely mentioned, it would be possible to understand what semantical statements say without thereby knowing what determinate directedness toward objective reality is said to be enabled by the expressions that the statements deal with, in linguistic acts in which those expressions occur; but that is not possible. In Sellars's view, the expressions that figure on the right-hand sides of these semantical statements are neither ordinarily used nor ordinarily mentioned, but *exhibit* their own propriety-governed use. We can understand this as a special kind of use, differentiated from using words in general in that it does not serve, for instance, to refer to an object; or alternatively as a special kind of mention, differentiated from mentioning words in general in that there is a presupposition that it is addressed to people who understand the expressions mentioned.[5]

---

[4] For 'means' in this role, see "Empiricism and the Philosophy of Mind," §31, p. 292 (163); for 'stands for', see *Science and Metaphysics*, chapter 3.

How can a statement that relates an expression only to another expression serve to determine an intentional character associated with the first expression, a role it plays in enabling linguistic acts it occurs in to be determinately directed toward elements in extralinguistic reality? As I said, the expression on the right-hand side of a specification of significance is supposed to exhibit its own propriety-governed use. If we were to formulate the relevant proprieties explicitly, we would be saying that there *ought to be* certain relations between, on the one hand, uses of the expression in question, considered now as elements in the real order (not the conceptual order), and, on the other hand, other elements in the real order, which can be, unlike the first set of parties to these relations, entirely extralinguistic.[6] That is, these proprieties require a certain determinate relatedness to extralinguistic reality on the part of occurrences in the real order in which speakers make ordinary uses of the expression that figures on the right-hand side of a semantical statement. By virtue of the nonordinary use to which the expression is put there, the content of that requirement—a certain determinate relatedness to extralinguistic reality—is reflected into what the statement says about the expression mentioned on its left-hand side, even though the statement relates that expression only to another expression. This is how a statement that affirms a relation between expressions is supposed to be able to capture the contribution made by the expression mentioned on its left-hand side to the intentional character (the directedness toward extralinguistic reality) of linguistic acts in which the expression figures.

---

[5] For a special kind of use, see, for example, "Empiricism and the Philosophy of Mind," §31, p. 292 (163): "not *mentioned* but *used*—used in a unique way; *exhibited*, so to speak." For a special kind of mention, see, for example, "Being and Being Known" (in *Science, Perception and Reality*), p. 55, where Sellars initially says that "'Mensch' signifies [another variant for "means" or "stands for"] *man*" "says, in effect, that the German word 'Mensch' has the same use as the English word 'man'"; then notes that on that account one could know what the statement of significance says without thereby knowing what 'Mensch' means; and finally suggests that "this can be remedied by interpreting [the statement of significance] as presupposing that the word 'man' is in the hearer's vocabulary," so that it amounts to "'Mensch' (in German) has the same use as *your* word 'man'."

[6] The linguistic terms of these relations that there ought to be are what Sellars calls "natural-linguistic objects." See "Truth and 'Correspondence'," in *Science, Perception and Reality*, p. 212: "although we may, indeed must, know that these linguistic objects are subject to rules and principles—are fraught with 'ought'—we abstract from this knowledge in considering them as objects in the natural order. Let me introduce the term 'natural-linguistic object' to refer to linguistic objects thus considered."

Sellars instructs us to model nonovert conceptual episodes on linguistic acts, and the structure of this account of meaning carries over into how he conceives the intentional character ("aboutness") of unexpressed thought. A conceptual episode's being intentionally directed toward an element in the real order is analogous to, say, a linguistic episode's containing an expression that functions as a name of an element in the real order. According to Sellars, this "aboutness" must not be conceived as a relation between an element in the conceptual order and an element in the real order. There are no semantical relations between the orders; aboutness, like meaning, is nonrelational. But the content of the relation, wholly within the conceptual order, that is affirmed by a statement of aboutness is partly constituted by relations that would ideally relate conceptual episodes considered as elements in the real order to other elements in the real order.

IV

Why does Sellars think we must thus explain away the appearance that there are semantical relations between the conceptual order and the real order? His remarkable paper "Being and Being Known" suggests an answer.

Sellars there aims to motivate his nonrelational conception of intentionality by discussing two versions of the opposing idea that intentionality can be relational. He formulates the opposing idea like this: "intellectual acts"—the "conceptual episodes" of "Empiricism and the Philosophy of Mind"—"differ *not* in their intrinsic character as acts, but by virtue of being directly related to different relata."[7]

---

[7] *Science, Perception and Reality*, p. 41. The terminology of "intellectual acts" reflects the fact that "Being and Being Known" is a discussion of how to interpret Aquinas. But the issues are exactly those raised by "Empiricism and the Philosophy of Mind" and *Science and Metaphysics*.

He finds the first version of this conception in an inheritance, on the part of Cartesian philosophy, of some scholastic apparatus. "For the Cartesian," Sellars says, "the immediate relatum is an item having being-for-mind ('objective' reality). Thus the thought of a golden mountain is a thought which is related to a golden mountain *qua* having being-for-mind, being for the mind that is thinking of it."[8] 'Content' is a natural label for something seen as having "objective" being *in* a thought. Using the term in that way, Sellars suggests that "the Cartesians postulated a domain of contents to mediate between the intellect and the real order" (*ibid.*, p. 42).

The second version Sellars identifies as the position of "the extreme realists of the early decades of the present century" (*ibid.*). In this version, the immediate relata of conceptual episodes are elements of the real order, not contents that mediate between conceptual episodes and elements of the real order. Philosophers in the Cartesian tradition had thought they needed those mediating contents in order to handle cases where the supposed real-order target of a conceptual episode does not exist; instead, in this version of the relational conception, the real order is expanded so as to include (for instance, as "subsistent") the nonexistent objects "which had puzzled previous philosophers into the theory of contents" (*ibid.*).

Sellars urges, reasonably enough given how he sees them, that each of these options has drawbacks. The mediating function of "contents" becomes problematic, as the Cartesian tradition succumbs to a tendency toward skepticism and idealism; and in any case the language of 'objective' being that Descartes inherits from scholastic philosophy merely labels the supposed relation between thought and objects, and does nothing toward explaining it. But the only alternative so far in view, extreme realism, is unsatisfactory just because of its extremism.

What seems to force a choice between these unattractive options is the assumption that conceptual episodes differ only extrinsically, in being related to different objects. This brings the assumption into doubt; and Sellars says, surely correctly, that the assumption—"the notion that acts of the intellect are intrinsically alike regardless of what they are about" (*ibid.*)—is odd anyway, even apart from the awkward choice it confronts us with.

At this point, Sellars writes:

> But what is the alternative? In general terms it is to hold that acts of the intellect differ intrinsically *qua* acts in a way which systematically corresponds to what they are about, i.e. their subject-matter (*ibid.*, p. 43).

---

[8] *Science, Perception and Reality*, p. 41.

This "alternative" is Sellars's nonrelational view of meaning and aboutness. Acts of the intellect—conceptual episodes—differ intrinsically, qua acts, in their intentional character. If we were allowed to understand intentional character as consisting in a relation to the extraconceptual order, we could put that thesis—that intentional character is the intrinsically differentiating character of a conceptual episode—by saying that conceptual episodes differ intrinsically *in* being about what they are about. But according to Sellars, intentional character is nonrelational. So the intrinsic differences between conceptual episodes cannot be differences *in* their subject matter; they can have to do with subject matter at all only by *systematically corresponding* to differences in subject matter. The systematic correspondence is the reflection into statements of significance—which, Sellars insists, set up relations only between elements in the conceptual order, not across the boundary between the conceptual order and the real order—of relations ideally required, relations there ought to be, between conceptual episodes considered as elements in the real order and other elements in the real order.

Here, then, Sellars recommends his position by eliminating alternatives. But as my paraphrase of the argument has brought out, there is a possibility he does not consider, even to reject it: that conceptual episodes might differ intrinsically, not in a way that *systematically corresponds* to what they are about, but *in* being about whatever they are about. He assumes that anyone who wants to say intellectual acts differ only in what they are about, as opposed to differing in some way that systematically corresponds to differences in what they are about, will admit to supposing that intellectual acts do not differ intrinsically at all. (That was his formulation of the alternative to his nonrelational conception.) He simply does not consider that someone might want to say a difference in what they are directed toward can itself be an intrinsic difference in intellectual acts. (That is just how it is with intuitions on the conception I have extracted from Kant's remark in the "Clue.") So the argument from exhaustion of possibilities is inconclusive; it depends on a tendentious formulation of the alternative to Sellars's position, according to which the alternative includes not just the idea that intentional directedness is relational but also—what Sellars assumes—that being related in the relevant way to an extraconceptual item cannot be intrinsic to an intellectual act.

And it is not just the last step of the argument that looks unsatisfactory in this light. Sellars's incomplete conception of the possibilities shapes his reading of the two other positions he considers. The prospects for positions on those lines look different once we query Sellars's assumption.

Consider again the apparatus Descartes inherits from scholastic philosophy. Crediting intellectual acts with "content" is, as Sellars sees, just another wording for the idea that elements of the real order have "objective" being in them. But that in turn can be, as Sellars does not see, just another wording for the idea that they have an intrinsic character that consists in their being immediately related to elements of the real order. So conceived, "content" does not *mediate* between the intellect and the real order, as in Sellars's reading. Rather, crediting intellectual acts with "content" is a way to express the thought that goes missing in Sellars's argument: that an unmediated relatedness to elements in the real order can be an intrinsic character of an intellectual act.

It is true that the scholastic terminology does no more than label the relatedness in question. But the apparatus need not preclude a vindicating explanation—perhaps on the lines of the one I am suggesting we can find in Kant. (This would require that when we set about entitling ourselves to the scholastic terminology, we focus primarily on intuitions, in which real objects are actually present to subjects, rather than on such intellectual acts as the thought of a golden mountain.) It need not be the apparatus itself that drives Cartesian reflection toward skepticism and idealism.

I think a Kantian rehabilitation of the scholastic-Cartesian apparatus would point to a version, purged of its extremism, of the twentieth-century realism that is the other option Sellars considers. In discussing this option, Sellars strangely implies that Russell's theory of descriptions belongs with a distinction between existence and subsistence, as another device for including nonexistent objects in the real order, where they can be relata for intellectual acts.[9] But surely the point of the theory of descriptions is exactly to avoid an apparent need for nonexistent real objects as relata for intellectual acts. Where a relational conception of intellectual acts would require them to stand in relations to possibly nonexistent objects, Russell instead takes their content to include *specifications* of objects. If no objects answer to the specifications, that does not threaten the contentfulness of the acts. There are indeed cases for which this Russellian strategy is unsatisfying, cases for which one wants to keep the idea of a relational intentionality in play even though the putative object of such intentionality may not exist. But one beauty of the idea that there can be illusions of relationally intentional content is that it enables us to gratify this quasi-Meinongian motivation without needing to postulate relations to merely subsistent objects. An illusion of a relation to an ordinarily real object does the work that seemed to require an actual relation to a merely subsistent object.[10]

---

[9] *Science, Perception and Reality*, p. 42: "Thus non-existent objects...found their place in the real order by means of a distinction between existence and subsistence and such other devices as Russell's theory of descriptions."

V

I have brought out how Sellars's attempt to justify his nonrelational conception of intentionality turns on the assumption that intentionality could be relational only if how a conceptual episode is directed at objective reality were not intrinsic to it. Reasonably enough, Sellars is not tempted by the idea that conceptual episodes are intrinsically alike, whatever they are about.[11] But so far from justifying the assumption that a relational conception of intentionality would commit us to that idea, Sellars does not even identify it as an assumption. This looks like a blind spot.

We can describe the blind spot in another way. In "Being and Being Known," Sellars formulates his conception of intentionality in terms of the Thomistic notion of the intellectual word. In this terminology, the intrinsic character of intellectual acts consists in their being (second) actualizations of intellectual words, and their intentionality is determined by the semantics of the intellectual words in question. Sellars reads this as a version of his own standard conception, according to which unexpressed thought is to be understood on the model of linguistic acts, literally so called—performances of overt speech. And his account of the semantics of intellectual words fits the familiar pattern. A statement of the meaning of an intellectual word affirms a relation only within the intellectual order, not between the intellectual order and the real order, but nevertheless contrives to capture an intentional directedness toward the real order, because what it says is partly constituted by an ideally required relatedness between elements in the real order. So we can formulate the possibility Sellars does not consider in terms of the semantics of intellectual words. The missing possibility is that a statement of the semantical character of an intellectual word might relate the word, as an element in the intellectual order, to an element in the real order, toward which intellectual acts in which the word figures are intentionally directed. I shall put this by saying that the semantics of the language of the intellect might be Tarskian.

---

[10] Sellars also ignores the difference between Russell's earlier conception of judgment as a relation to a state of affairs, which may or may not be existent, and his later conception of judgment as a multiple relation to objects and properties Here, again, the point of the later conception (unsatisfactory as it certainly is) is to avoid the Meinongian commitment that the earlier conception incurs, to nonexistent states of affairs as real relata for false judgments.

[11] Sellars is happy to "take for granted that the concept of a diaphanous act...is unsatisfactory" (*Science and Metaphysics*, p. 63). The concept of a diaphanous act is the concept Sellars finds in the views to which, in "Being and Being Known," he represents his own position as "the alternative": the concept of an act that differs from other acts in what it is intentionally directed toward, but not intrinsically. See *Science and Metaphysics*, p. 34.

This would be a counterpart to the thought that the semantics of ordinary words might be Tarskian, which is ruled out by Sellars's doctrine that meaning is nonrelational. He occasionally discusses Tarskian semantics for ordinary words, but his discussions are quite unsatisfactory, and we can see this as a symptom of the same blind spot.

Sometimes he suggests that the very idea of word-world relations as they figure in Tarskian semantics is "Augustinian," in the sense that fits the opening sections of Wittgenstein's *Philosophical Investigations*.[12] But this is simply wrong. It is perfectly congenial to Tarskian semantics to say that the notions of such word-world relations as denotation and satisfaction are intelligible only in terms of how they contribute to capturing the possibilities for "making moves in the language-game" by uttering whole sentences in which the relevant words occur. These relations between words and elements in the extralinguistic order should not be conceived as independently available building blocks out of which we could construct an account of how language enables us to express thoughts at all.[13]

In other places, Sellars suggests that proponents of relational semantics conceive the word-world relations that they take semantical statements to affirm in terms of "ideal semantical uniformities." This is an allusion to those propriety-governed genuine relations, between linguistic acts considered as elements in the real order and other elements in the real order, that figure in his picture as partly constitutive of the content of semantical statements.[14] Here Sellars is reading Tarskian semantics in the light of his own understanding of the possibilities. Statements of those "ideal semantical uniformities"—which are not themselves semantical statements, though they enter into the determining of what semantical statements say—are the closest his view can come to the idea of statements that are both semantical and deal with relations to elements in the real order. So Sellars takes it that proponents of relational semantics mistakenly think these statements of "ideal semantical uniformities" *are* semantical statements. This is to assume that his opponents are working within a dimly grasped version of his structure, and misconstrue the significance of its elements.[15]

---

[12] G.E.M. Anscombe, trans. (Oxford: Blackwell, 1951). For a suggestion in this direction on Sellars's part, see "Empiricism and the Philosophy of Mind," §30, pp 290-91 (161-62). Compare Brandom's contemptuous remarks about "a supposed word-world relation of reference"—*Making it Explicit*, pp. 323-25.

[13] A way of putting this is to say that Tarskian semantics can perfectly well accommodate the point Sellars makes about truth, directing it against Carnap—*Science and Metaphysics*, pp. 100-02. (Carnap tends to be teamed with Tarski in Sellars's discussions of semantics; see, for example, *Science and Metaphysics*, p. 83: "semantics of the Carnap-Tarski type.") Sellars's point does not tell against doing semantics in terms of relations; rather, it is a way of saying why the derivability of conclusions conforming to "Convention T" is a good adequacy condition on a relational semantical account of a language. For a clear statement of the point that there is nothing "Augustinian" about Tarskian semantics, see Donald Davidson, "In Defense of Convention T," in his *Inquiries into Truth and Interpretation* (New York: Oxford, 1984).

But that is not what is meant by saying that statements of, for instance, the form "...denotes—" relate words to objects. The point is rather this. First, the expression that figures on the right-hand side of such a statement is used in an ordinary way, not in the peculiar way that figures in Sellars's account of semantical statements. So we can see the statement as itself affirming a relation between the expression mentioned on the left-hand side and whatever element in the real order can be mentioned by a standard use of the expression on the right. Second, and crucially, the ideality or normativity that is relevant to such a statement is not that of the "ideal semantical uniformities" that figure in Sellars's own picture of the semantical. 'Denotes' expresses a relation between elements of the linguistic order and elements of the extralinguistic order, the very idea of which is—to borrow a Sellarsian phrase—fraught with 'ought',[16] in a way that reflects what ensures that this conception of semantics is not "Augustinian." The very idea of such relations makes sense only in the context of how they enter into determining the conditions under which whole sentences are *correctly* or *incorrectly* asserted. The normativity expressed by those uses of 'correctly' and 'incorrectly' is reflected back into the content of such concepts as that of denotation.[17]

---

[14] See *Science and Metaphysics*, pp. 86-87; the same suggestion is made at p. 112. The "uniformities" in question here are those described at pp. 75-77

[15] See *Science and Metaphysics*, p. 86: "It is a mistake to suppose, as Carnap does, that semantical statements in *his* sense, i.e. statements which involve such expressions as 'denotes' or 'designates', are semantical statements in the sense that they formulate (ideal) semantical uniformities."

[16] See, for example, "Truth and Correspondence," in *Science, Perception and Reality*, p. 212.

Sellars holds that we should understand nonovert conceptual episodes on the model of overt linguistic acts. So if it were acceptable to understand the semantics of intellectual words in Tarskian terms, that would be because our analogical understanding of intellectual words can exploit its being acceptable to understand the semantics of ordinary words in Tarskian terms. But Sellars's blind spot occludes the idea that the semantics of intellectual words might be Tarskian, except as an expression of the unprepossessing conception in which intellectual acts have no intrinsic character. This helps to make it intelligible that a relational conception of the semantics of ordinary words does not come into Sellars's thinking in the authentic form in which I have just sketched it.[18]

VI

At the beginning of these lectures, I spoke of a structural feature of Sellars's thinking that, I suggested, makes a fully Kantian vision of intentionality inaccessible to him. I have now come close to uncovering what I meant. Sellars cannot see how a determinate intentional directedness can be both a relation to an element in the real order and an intrinsic character of a conceptual occurrence, and this corresponds to an inability to see how denoting, say, can be a relation that relates elements of the conceptual order to elements of the real order.

Why does Sellars not contemplate the possibility I am urging?[19] I do not mean to suggest that his blind spot is superficial.

---

[17] Another consideration Sellars brings to bear against "semantics of the Carnap-Tarski style" connects this region of his thinking with his doctrine about the manifest image and the scientific image; he objects to a relational semantics for the language of the manifest image, on the ground that it would commit one to the reality of the objects of the manifest image. See "The Language of Theories," in *Science, Perception and Reality*, p. 109, note 3.

[18] This has a damaging effect on the thinking of some people who have been influenced by Sellars. Rorty, for instance, sees that the Tarskian semantics Davidson envisages for natural languages involves word-world relations, and concludes that Davidsonian semantical talk, so far from being fraught with 'ought', is not even coherently combinable with a way of talking about language that is fraught with 'ought'. See "Pragmatism, Davidson and Truth," in Rorty's *Objectivity, Relativism, and Truth* (New York· Cambridge, 1991). This misses the fact that the very idea of the word-world relations in question (for instance, denotation) is itself already 'ought'-laden, in a different way from any that comes into view in Sellarsian semantics. Rorty's reading of Davidson looks like a descendant of Sellars's blind spot.

[19] It is not that Sellars cannot see how something as natural as sensory consciousness could be shaped by conceptual capacities, our possession of which is what elevates us above mere nature. (That is the supposed difficulty I tried to deal with in *Mind and World*.) Sellars has, and exploits, the idea of conceptual shapings of sensory consciousness.

It helps to compare two ways of interpreting the idea that our thought of meaning and aboutness is fraught with 'ought'.

On Sellars's interpretation, the content of a statement of significance is a reflection, into a statement of a relation within the conceptual order, of relations that there ought to be, according to the proprieties that constitute a linguistic practice, between two sets of elements in the real order, one of which comprises linguistic items considered in abstraction from the practical proprieties in virtue of which they are meaningful at all. The 'ought' with which meaning and aboutness are fraught gets into the picture as a sentential operator, in whose scope there occur specifications of relations that would ideally hold between linguistic items so considered and other elements in the real order. The content of the 'ought' with which some fact about significance is fraught—what it is that, according to the 'ought' in question, ought to be the case—can be factored out from the statement of significance and specified in terms that are not themselves meaning involving.[20]

Contrast the Tarskian conception. Here denoting, say, is a relation, itself fraught with 'ought', between an element in the linguistic or conceptual order and an element in the real order. But in this conception it is not the case that to a statement of what some expression denotes there corresponds an 'ought' whose content could be factored out, so that the 'ought' in question could be seen as a separable determinant of the semantical character of the expression. The 'ought' with which a statement of what some expression denotes is fraught reflects how the expression's denoting what it denotes enters into determining conditions for the correctness of assertions. The normativity that the 'ought' conveys is an aspect of the expression's meaning, its impact on the significance of sentences it occurs in. This normativity is not something presemantical that could figure in constitutively explaining what meaning is from outside the semantical.

So I can formulate a structural feature of Sellars's thinking like this: he cannot see how semantical thinking could be 'ought'-laden except by taking it to be fraught with 'oughts' that can be seen as determining significance from outside the semantical. He thinks there must be proprieties of (what is in fact) linguistic practice, formulable in nonsemantical terms, that underlie and constitute the semanticity of linguistic expressions.

---

[20] See, for example, p. 92 of *Naturalism and Ontology* (Reseda, CA: Ridgeview, 1979): "The rule governed uniformities...which constitute a language (including our own) can, in principle, be exhaustively described without the use of meaning statements."

Kant suggests an understanding of thought's having objective purport that centers on the immediate presentness of objects to conceptual consciousness in intuition. I have suggested a reading in which this immediate presentness is relational. Sellars wants to exploit the Kantian idea of a transcendental role for sensibility, but for him the idea that intuitional content might be understood in terms of a relation between the conceptual order and the real order is not an option. Now the contrast I have drawn between two ways of understanding fraughtness with 'ought' suggests that we can see this as reflecting a general conception of what it takes to execute Kant's transcendental project. For Sellars, our entitlement to see elements in the conceptual order as intentionally directed toward elements in the real order has to be transcendentally secured from outside the semantical, from outside the conceptual order. We have seen this structural feature of Sellars's thinking operative in connection with the 'ought' with which our thought of meaning and aboutness is fraught, but the idea that the transcendental work needs to be done from outside the semantical has a more general application. Given a conviction to that effect, a transcendental role for sensibility can only be the sort of thing Sellars envisages, a matter of conceptual activity being guided by "sheer receptivity." On this view, we cannot spell out a transcendental role for sensibility in terms of the immediate presence of objects to intuitionally structured consciousness, as in the reading of Kant I have recommended. That would be already a case of conceptual directedness toward the real, so it could not figure in a vindication, from outside, of the very idea of conceptual directedness toward the real.

How can there be a semantical relatedness, itself fraught with 'ought', between the conceptual order and the real order? It might be the beginning of an answer to say: this 'ought'-laden relatedness to the real order must be itself embraced within the conceptual order. I have deliberately given that formulation a Hegelian ring, in order to suggest that what I have been representing as Sellars's blind spot, his inability to contemplate the possibility that intentionality might be relational, is part of a package with his conviction that to give philosophical reflection about intentionality a Hegelian shape is to abandon objectivity rather than to vindicate it.

We now have a variety of ways of describing this structural feature of Sellars's thinking. One is that it reflects his conviction that Hegel merely muddies the Kantian waters.[21] Against this, I have urged that, if we see intuition in the way Kant proposes, we can take perceived objects themselves to supply the external constraint on conceptual goings-on for which Sellars thinks we need to appeal to "sheer receptivity." A nonsheer receptivity is operative in intuition so conceived, and that is enough to undermine the threat of "idealism," the threat that the supposed objects of these conceptual shapings of consciousness can only be projections of our conceptual activity. At any rate there is nothing against this reading of the transcendental role of sensibility except the putative reasons yielded by scientism for denying genuine reality to the objects that, speaking within the manifest image, we say are immediately present to us in intuition, and hence for refusing to allow that those objects themselves provide the transcendentally needed external constraint on conceptual activity. Discounting such scientism, we can refuse to be frightened by Sellars's invocation of Hegel. I am encouraged in this by the availability of another way to describe Sellars's idea that there must be guidance by "sheer receptivity": it reflects his blind spot for the Tarskian approach to the semantical.

---

[21] See *Science and Metaphysics*, p. 29.

It may seem a dizzying project to embrace relatedness to the real order within the conceptual order. But Tarskian semantics points to a sober interpretation. It is crucial to a proper understanding of Tarskian semantics that we have to *use* the words on the right-hand sides of semantical statements. (Contrast Sellars's conception of the way words figure on the right-hand sides of statements of meaning.) Sellars himself holds that unexpressed thought is to be understood on the model of linguistic acts. Exploiting that, we can apply the same point to unexpressed thought: we have to *use* the words that figure in specifications of what nonovert conceptual episodes are intentionally directed toward. In statements of meaning and aboutness, we relate the conceptual order to the real order, mentioning elements of the real order by making ordinary uses of the words on the right-hand sides of these statements. But we affirm these relations without moving outside the conceptual order—without doing more than employing our conceptual capacities.

How can our minds get into relations to elements in the real order simply by acquiring suitably shaped conceptual contents? That is just what we are enabled to find unmysterious by putting intuitional content, understood in the way Kant indicates, at the center of our picture of the conceptual.

VII

I began these lectures by suggesting that Sellars's thinking is especially instructive toward understanding Kant, and thereby finding intentionality unproblematic. I can now end by giving that suggestion a somewhat more determinate shape. The point is that the key question for understanding Kant, and thereby seeing how to become comfortable with intentionality, is just the question Sellars brings into focus: Can the transcendental project be acceptably executed from within the conceptual order, or does it require a sideways-on point of view on the directedness of the conceptual at the real?

In Lecture I, I distinguished transcendental philosophy as such from a conception of it as requiring the sideways-on perspective. Sellars's position exemplifies this more specific conception, according to which the transcendental project needs to be undertaken from outside the thought whose objective purport it seeks to vindicate. Correctly in my view, Sellars fails to find anything answering to this conception of transcendental philosophy in Kant's "Aesthetic"; and he contrives to find this kind of thing adumbrated in the "Analytic" only by ignoring Kant's insistence that appearances, things as they appear, are the same things that can also be conceived as things in themselves. The very idea of transcendental philosophy is a Kantian invention, and we can hardly suppose there really is none in the first *Critique*. The fate of Sellars's reading suggests that the thought Hegel tries to capture with the image of Reason as subject to no external constraint—the rejection of a sideways-on standpoint for philosophy—is already Kant's own thought.

It is common for people not to see so much as a possibility that intentional directedness might be a relation to objects; this shows in the resistance or incomprehension typically encountered by Evans's work. Often this is combined with taking for granted the idea that conceptual activity is intentionally directed toward the world. Now Sellars is special in being responsive to the Kantian thought that we need a transcendental exercise in which we show our entitlement to the very idea of objective purport. He is also responsive to the more specific Kantian thought that this transcendental exercise must center on intuition, so that we can exploit the receptivity of sensibility. But he is unresponsive to the Hegelian conceit of incorporating receptivity within Reason, and I have tried to display this as a blindness to a more soberly describable possibility. Given his conviction that the transcendental exercise must be undertaken from outside the conceptual, Sellars's responsiveness to Kant gives him no alternative but to construe the transcendental role of sensibility in terms of guidance by "sheer receptivity."

Suppose we agree with Sellars that it is an insight on Kant's part that the receptivity of sensibility must play a transcendental role. Reflecting on the context Sellars supplies for that thought, we can see that, aside from minor details, the conviction that Kant is right about the significance of sensibility presents us with a quite simple choice: either Sellars's picture of guidance by "sheer receptivity," or the idea I have recommended, that the guidance Sellars thinks we need to credit to "sheer receptivity" can be displayed, in the course of the transcendental project, as exercised by the immediate objects of perception themselves.[22] I hope to have made it plausible that there is more to be said for the second option than Sellars allows.

JOHN MCDOWELL

University of Pittsburgh

---

[22] The only alternative, at this level of generality, is to deny a transcendental role to sensibility altogether. Brandom's *Making It Explicit* is the most worked-out attempt I know to take this non-Kantian path and still purport to accommodate intentionality. In chapter 4, where Brandom undertakes to deal with empirical content, he deliberately refrains from attributing a transcendental role to sensory consciousness, thus denying himself the resources for a Kantian notion of intuitions. (The same fact about Brandom's thinking is reflected in his noncentral placing of the situations that permit "strong de re ascriptions," in chapter 8.) In his chapter 8, Brandom undertakes to provide for the "representational character" of thought and language, without a transcendental exploitation of anything on the lines of intuitions, by appealing to the idea of a specification by *A* of what *B*'s thought is about in which *A* takes responsibility for the way the object is specified. In "Replies," in *Philosophy and Phenomenological Research*, LVII (1997). 189-204, he claims that this apparatus legitimates an idea of word-world relations. I think this cannot work. See *Science and Metaphysics*, pp. 82-87, where Sellars shows, I think, that, if we start with the idea that word-world relations are impossible or at least problematic (as Brandom does), then considerations about substitution inferences, of the sort Brandom exploits, cannot get us any closer to vindicating semantical relations between the conceptual order and the real order.

MEMO TO AUTHORS

The editors cannot afford to send to the printer manuscripts that are difficult for the compositor to read or footnotes that need to be completed in proof. They ask that manuscripts submitted be prepared according to the following guidelines:

(1) All material, including the quotations and footnotes—especially footnotes—must be typed, double-spaced, on one side of the paper, with ample margins left and right, top and bottom.
(2) No bibliographies will be published; bibliographical material should be put into the footnotes.
(3) Footnotes should be complete, according to Journal style (see any issue). Journal citations should include name of journal in full, volume number, issue number, month or season, year, and pages. Book references should include publisher, city of publication, and date. Footnotes should be gathered at the end of the article.
(4) Quotations should be checked by the author; capitalization, italics, spelling, and the like should match the original. Corrections should not have to be made (expensively) in proof.
(5) If there are diagrams, they should be provided camera-ready by the author. If there are special symbols, they should be widely spaced and clearly marked; if there are many, a list should be provided. Standard logic symbols are available, and letters of the Greek and script alphabets; no boldface.
(6) Computer copy can be extremely difficult to read; only letter-quality is acceptable. It should be double-spaced and not justified. The print should be dark enough and the letters widely enough spaced so that both letters and punctuation marks can be read easily.

The editors are willing to read and evaluate manuscripts that do not meet these standards, provided they are genuinely legible and not unreasonably long. They may ask for revision of excessively long manuscripts, and they require reformatting of accepted manuscripts, if necessary, before scheduling for publication.

The *Journal of Philosophy* is published monthly, in issues that (ordinarily) contain only 56 pages of copy; this limits the length of papers the editors can normally publish to no more than 7,500 words, or eighteen Journal pages. A more precise way of determining length is by characters: 2,850 characters per Journal page, or 51,300 characters per Journal article of 18 pages.

The *Journal of Philosophy* does not publish papers that are primarily historical or expository; the editors urge that such manuscripts be submitted elsewhere.

THE JOURNAL OF PHILOSOPHY

VOLUME XCV, NO. 10, OCTOBER 1998

PROOF AND TRUTH: THROUGH THICK AND THIN*

In Jill Paton Walsh's novel *The Knowledge of Angels*,[1] Palinor, a principled agnostic, carries on an extended disputation with Beneditx, a scholarly monk. The issue is whether the existence of God can be known by reasoning alone—or whether it can be known at all. After Palinor rebuts Anselm's ontological argument, Beneditx argues that there is such a thing as truth:

> ...if there is anything true, then there must be truth. God is truth itself. All discourse, therefore, all statements in all sciences, all refutations and disproofs contain the idea of God, because they contain the idea of truth. The existence of God is self-evident, everywhere assumed, even when men believe themselves to be talking of other things (*ibid.*, pp. 117-18).

Palinor points out a flaw in this reasoning: "I would say rather, if there is anything true, there must be *truths*. A very different statement from yours." Beneditx says that "for a statement to be true, it must partake of the truth." Palinor retorts as follows:

> I would not put it that way, Beneditx, and some of our difficulty in agreeing arises from the fact that we put things differently, and how you put things makes a considerable difference to how they seem.... I would not think of truth as single, like a great ocean, but as multiple, like many

---

* This paper was provoked by a talk that Paul Horwich gave at the University of St. Andrews in the Spring of 1996, and by discussions with him. Many thanks to Neil Tennant, Robert Kraut, Sven Rosenkranz, Adam Cook, Thomas Hofweber, Alan Weir, Penelope Maddy, Mark Wilson, Tim Carlson, Harvey Friedman, Anil Gupta, Michael Dunn, and Crispin Wright for conversations on this topic and, in particular, for helping me through some of the literature on deflationism. I also thank the Logic and Metaphysics Research Seminar at the University of St. Andrews, the Indiana Logic Colloquium, and the Logic, Language, and Science Colloquium at Ohio State University for devoting fruitful sessions to this project.

[1] Boston: Houghton Mifflin, 1994.

0022-362X/98/9501/493–521

> rivers.... If I say to you that the bird singing on the branch there is a warbler, how would you discover the truth of my statement? The truth of a statement that names something can be confirmed in a book of names, or by asking a speaker of the language in which the name is uttered. If I tell you that my servant can swim, the truth of the statement can be discovered by throwing him in the river; if I tell you that in my country the sun sets at midnight on midsummer day you will need to travel there at the right season to confirm it, and so on. Each of these statements is true in a different way, and a different process of confirmation—or, of course, refutation—is required for each, although the words "true" and "false" are used to describe the results of all the investigations. In the many different kinds of procedure needed to verify things, I would find evidence that truth is of many kinds: were there such a thing as "The Truth," surely one way of investigating matters would always reveal it. You say that God is truth itself—could he be said to be truths (*ibid.*, pp. 118-19)?

Beneditx concedes that this last "doesn't sound right at all, for God is one and unchanging." Palinor continues:

> You see what a difference it makes when one uses a different way of speaking.... The existence of truths does not so easily lead one to say that there is one overriding truth for God to be. Of course, if you can lead me from truths to one truth, I will gladly follow you. All you will then need to do is to prove the identity of this one truth with the God of your belief (*ibid.*, p. 119).

Palinor thus wins the opening rounds, and he goes on to a tragic victory. Beneditx is robbed of his faith, and Palinor is burned as a heretic.

If we strip out Palinor's verificationism, we see that he was well on the way to deflationism toward truth, including the ubiquitous deflationary phrase: 'and so on'. Palinor may very well have gotten to deflationism if he felt some need for a theory of truth and was not distracted by the other arguments for the existence of God. The main theme of deflationism is that truth is not a metaphysically substantial property, and it has no nature. My purpose here is to examine the extent to which the deflationist's claim can be squared with the logician's use of truth as a sign of expressive and proof-theoretic *strength*.

We know, from Tarski's theorem, that, if a language $\mathscr{L}$ can express a small amount of arithmetic, then no consistent theory in $\mathscr{L}$ can express a truth predicate for $\mathscr{L}$. Kurt Gödel and Alfred Tarski both discovered that one can simulate self-reference in a language of arithmetic. Thus, if truth could be defined, then there would be a liar sentence:

> This sentence is not true.

which would be inconsistent.

Suppose, then, that a logician is dealing with two theories T_1 and T_2, possibly using different languages. Suppose that in T_1 we can characterize the syntax of T_2 and define a truth predicate for T_2. That is, suppose that there is a formula $\Phi(x)$ in T_1 such that for each name s, of each sentence Ψ in the language of T_2, we have that $\Phi(s)$ is true in some (intended) interpretation of T_1 if and only if Ψ is true in some (intended) interpretation of T_2. This is the T-scheme. It follows from Tarski's theorem that T_1 is stronger than T_2. In a sharp sense, we can express more, and prove more, in T_1 than we can in T_2. We see this *because* the one theory has a truth definition for the other.

I do not know if there is a consensus among deflationists over how metaphysical substance lines up with expressive and proof-theoretic power, but there is at least cause for thought here. Why is it that the metaphysically thin, or natureless, or lightweight concept of truth should be a sure-fire sign of expressive and proof-theoretic strength? I suppose that some deflationists can joyfully accept this situation, saying that metaphysical substance is irrelevant to anything expressive or proof theoretic. I propose to follow some examples to see what is needed in order to be so sanguine. My conclusion is that the deflationist needs a very strong, noneffective notion of logical consequence.

Contemporary writers articulate the deflationist theme in different ways. One common claim is that there is little more to the notion of truth than is given by the disquotation scheme. We can formulate this scheme as a pair of rules, an introduction rule and an elimination rule:

(TI) From a sentence Φ, infer Tp, where p is a canonical name of Φ.
(TE) From a sentence Tp, infer Φ, where p is a canonical name of Φ.

The issue turns on what truth is for. Why do languages have a truth predicate at all? Typically, deflationists argue that a truth predicate is only a linguistic device for indirect endorsement and part of a linguistic device for generalization. I do not have any other uses to propose, but let us see what is involved in these two.

Indirect endorsement is a device to save time and ink. We can endorse the Gödel sentence without writing it out, by saying 'The Gödel sentence is true'. If we think Karl is honest and reliable, we can confidently say 'What Karl said at the trial is true' without bothering to find out exactly what he did say at the trial. The disquotation scheme provides a license for indirect endorsement, and perhaps that much is semantically and metaphysically thin. Dorothy Grover, Joseph Camp, and Nuel Belnap call indirect endorsement the "lazy" use of "truth."[2]

---

[2] "A Prosentential Theory of Truth," *Philosophical Studies*, XXVII (1975): 73-125

Generalization is another story. We want to say things like 'Everything the rabbi says is true'. Since the range of the quantifier here is at least potentially infinite, we cannot infer the generalization from any finite number of instances of the disquotation scheme. We also cannot infer the generalization from infinitely many instances if our consequence relation is compact, like the consequence relation of first-order logic.[5] Therefore, every account of truth—deflationist or otherwise—must involve some device for generalizing; otherwise, some important aspect of the use 'truth' is left out. Most deflationists recognize this.

Let $\mathcal{L}$ be a language whose syntax is effective, and suppose that $\mathcal{L}$ can express a certain amount of arithmetic. For convenience, I stipulate that the variables of $\mathcal{L}$ range over natural numbers. We associate each sentence Ψ of $\mathcal{L}$ with a natural number $\ulcorner\Psi\urcorner$, its Gödel number. For each number n, let Φ_n be the sentence with Gödel-number n.

Assume that there is an effective, sound theory A in the language $\mathcal{L}$ and that A contains the rudimentary axioms of arithmetic. Then a "proof predicate" can be defined in the language. That is, there is a formula $PRF(x,y)$ such that for any numerals m,n, $PRF(m,n)$ is provable in A if and only if n denotes the Gödel number of a proof in A of Φ_m; and $\neg PRF(m,n)$ is provable in A if and only if n does not denote the Gödel number of a proof in A of Φ_m. Let G be a fixed point for $\neg\exists y PRF(x,y)$, so that

$$G \equiv \neg\exists y PRF(\ulcorner G\urcorner, y)$$

is provable in A. Roughly, Gödel's first incompleteness theorem is that G is true, but not provable in A (provided that A is consistent). One can show in A that G is equivalent to $\neg\exists y PRF(\ulcorner 0=1\urcorner, y)$, which we abbreviate *Con*. The sentence *Con* "says" that A is consistent. Since G is not provable, then neither is *Con*. This is Gödel's second incompleteness theorem.

Suppose, now, that we follow Tarski and add a truth predicate to $\mathcal{L}$. That is, we augment the language with a new predicate T, whose intended interpretation is as a truth predicate for the original language $\mathcal{L}$. We also add axioms to A sufficient to prove every T-sentence:

$$T(n) \equiv \Phi_n$$

---

[5] We can infer 'everything the rabbi says is true' from (TI), 'the rabbi says that Pork is tref', 'pork is tref', 'the rabbi says that carrots are kosher', 'carrots are kosher', and 'the rabbi does not say anything else'. We can also infer 'everything the rabbi says is true' from 'the rabbi does not say anything'. Clearly, we need the means to infer the generalization when the range is not so limited.

To accomplish this, we can simply add the T-sentences as new axioms or else follow Tarski[4] and introduce terminology for satisfaction (in which case we only need finitely many new axioms). Call the augmented language $\mathscr{L}'$ and the augmented theory A'. In the T-scheme, the (meta)variable n ranges over the Gödel numbers of sentences in the original language $\mathscr{L}$, not the augmented language $\mathscr{L}'$, and so we stay far removed from paradox. This may make the account incomplete, but not incorrect.

If the original theory A is first-order, then it can be arranged for the new system A' to be a conservative extension of A: if Φ is a sentence in the original language $\mathscr{L}$, then Φ is provable in A' only if Φ is provable in A. In other words, adding a truth predicate to the original theory need not allow us to prove anything in the original language that we could not prove before we added the truth predicate. This should be pleasing to the deflationist. If truth/satisfaction is not substantial—as the deflationist contends—then we should not need to invoke truth in order to establish any results not involving truth explicitly.

We can put the situation model theoretically. Let M be any model of A. Then the T-sentences determine an extension for the new predicate T, and with this extension M can be extended to a model M' of A' in the language $\mathscr{L}'$. Thus, any model for a theory *without* a truth predicate can be extended to a model *with* one. This is more grist for the deflationist mill that truth is metaphysically thin.

Let me elaborate the philosophical point. I submit that in one form or another, conservativeness is essential to deflationism. Suppose, for example, that Karl correctly holds a theory B in a language that cannot express truth. He adds a truth predicate to the language and extends B to a theory B' using only axioms essential to truth. Assume that B' is not conservative over B. Then there is a sentence Φ in the original language (so that Φ does not contain the truth predicate) such that Φ is a consequence of B' but not a consequence of B. That is, it is logically possible for the axioms of B to be true and yet Φ false, but it is not logically possible for the axioms of B' to be true and Φ false. This undermines the central deflationist theme that truth is insubstantial. Before Karl moved to B', $\neg\Phi$ was possible. The move from B to B' *added* semantic content sufficient to rule out the falsity of Φ.

---

[4] "Der Warheitsbegriff in dem formalisierten Sprachen," *Studia philosophica*, I (1933): 261-405, translated as "The Concept of Truth in Formalized Languages," in Tarski, *Logic, Semantics and Metamathematics* (New York: Oxford, 1956; second edition, edited by John Corcoran, Indianapolis: Hackett, 1983), pp. 152-278.

But by hypothesis, all that was added in B' were principles essential to truth. Thus, those principles have substantial semantic content.[5]

This little argument presupposes that the dreaded property of metaphysical robustness lines up with logical consequence or semantic content. The presupposition is that, if a set Γ of sentences not involving truth are jointly possible, but are incompatible with certain principles Δ of truth, then the truth principles Δ are robust. They tell us what is possible and not possible concerning the truth-free fragment of language.

To be sure, there is no consensus today on the interlocked notions of logical consequence, semantic content, metaphysical strength, and metaphysical possibility; and some philosophers are convinced that most of these notions are obscure and bankrupt. This might give the deflationist some room to maneuver, and it leaves the issue hard to adjudicate. Nevertheless, it seems that in some sense or other, the deflationist is committed to the conservativeness of truth. Deflationism presupposes that there is some sense of 'consequence' according to which truth is conservative.

Returning to the case at hand, we see that A' is both proof-theoretically and model-theoretically conservative over A. So far so good, but some (potentially) troubling news for the deflationist is on the horizon. All would be well for the deflationist if the augmented theory A' gave us just about everything we need concerning arithmetic truth. But it does not. As noted above, the T-sentences are sufficient to use the system for indirect endorsement. One can assert $T(n)$ without bothering to write out Φ_n, or even to find out exactly which sentence Φ_n is. We also need, however, to express and establish *generalizations* concerning truth. Since the original language $\mathscr{L}$ contains ordinary quantification, one can state in $\mathscr{L}'$ that all of the axioms of A are true and one can state in $\mathscr{L}'$ that the rules of inference preserve truth. Since these generalizations are obviously correct, an adequate theory of truth should have the resources to *establish* them. It follows, or should follow, that *all of the theorems are true.* In $\mathscr{L}'$ this is:

$$(*)\quad \forall x(\exists y \mathrm{PRF}(x,y) \rightarrow T(x))$$

Recall that $\mathrm{PRF}(x,y)$ is only a proof predicate for A, not A'. Thus, the sentence (*) says in $\mathscr{L}'$ that all of the theorems of A are true. If we could prove (*) in A', then since $\neg T(\ulcorner 0=1 \urcorner)$ is a consequence of one of the T-sentences, we would have that

$$\neg \exists y \mathrm{PRF}(\ulcorner 0=1 \urcorner, y)$$

---

[5] I am indebted to Crispin Wright for getting me to put the point this way.

is provable in A'. This last is the sentence *Con*, which does not contain the truth predicate, and so is in the original language $\mathscr{L}$. By conservativeness, we would have that *Con* (and so G) is provable in A, which contradicts Gödel's incompleteness theorem. Thus (*) is not provable in A'.

So, if our deflationist takes A' to capture just about everything essential to truth concerning $\mathscr{L}$ and if the consequence relation is first-order, then our deflationist cannot say that all of the theorems of A are true. This is a reductio ad absurdum on the conjunction of the antecedents of this conditional—that the consequence relation is first-order and A' captures just about everything essential to arithmetic truth. The theorems of A *are* true, and we can see that they are. To put it informally, once our subject has taken on the truth predicate, and he notices that all the axioms of A are true and that the rules of inference preserve truth, he concludes that every theorem of A is true. He also knows (from (TE)) that '0=1' is not true, and so '0=1' is not a theorem of A. So our subject concludes that A is consistent (*Con*) and that G is true. The defect of A' as a theory of arithmetic truth is that it cannot reproduce this simple, informal reasoning.

If A' invokes something like Tarskian satisfaction, then it has a natural extension that does contain *Con* and G as theorems. The extended theory flows from the intuitive notion of natural number. If A is first-order, then the arithmetic in A contains an induction scheme:

$$[\Phi(0) \mathbin{\&} \forall x(\Phi(x) \rightarrow \Phi(sx))] \rightarrow \forall x\Phi(x)$$

one instance for each formula $\Phi(x)$ *in the original language* $\mathscr{L}$. That is, we can apply induction to a formula Φ in A and thus in A' only if Φ does not contain any instances of the truth predicate T. Suppose that we expand A' to allow instances of the induction scheme in which the formula Φ contains T. That is, we expand the induction scheme to the extended language $\mathscr{L}'$. Call the resulting theory A''. In that theory, we can do an induction on the length of proofs and establish that all the theorems (of A) are true, along the lines of the above informal reasoning. Thus, the sentence (*) is provable in A''. *Con* follows as above, and so does the Gödel sentence G of the original theory A. Thus, A'' is not a conservative extension of A, and so there are some (nonstandard) models of A that cannot be extended to become models of A''.

In light of the conservativeness requirement, if the deflationist takes the consequence relation to be first-order, then he cannot accept A'' as even a partial account of arithmetic truth. If truth is admitted into the induction principle, it is counterintuitive to maintain that the notion is thin and insubstantial. How thin can the notion of arithmetic truth be, if by invoking it we can learn more about the natural numbers? By taking on this supposedly insubstantial, mere device for indirect endorsement, we can establish facts about the natural numbers that we could not establish before. We cannot be definitive yet, since the conclusion depends on how the deflationist notions of "insubstantial" and "natureless" are articulated, and it depends on the notion of logical consequence. Surely, however, there is some tension.

Michael Dummett[6] comes to a similar conclusion. Suppose that someone advances the deflationist "view that we can have no notion of truth for arithmetical statements...beyond a strictly disquotational one" (*ibid.*, p. 333). Dummett argues that on this view, "an intuitively correct formal system will be one...whose axioms we *treat* as true, but no reflection will guarantee that we have any reason to think it consistent, nor, hence, any to treat the undecidable statement as true" (*ibid.*). In other words, the strictly disquotational notion of truth does not allow the inclusion of formulas with the truth predicate in the induction scheme, and does not support the relevant generalization (*). If the consequence relation is first-order, then the notion of truth (for $\mathscr{L}$) in A'' is more substantial than the mere disquotational one, and so this notion is not available to the deflationist.

On the other hand, whether one is a deflationist or not, there is no good reason to demur from the extension of the induction scheme to the new language. There is no reason to demur from A''. Informally, the induction principle is that for any well-defined property (or predicate), if it holds of 0 and is closed under the successor function, then it holds of all natural numbers. It does not matter if the property can be characterized in the original, first-order theory. As Dummett put it: "It is part of the concept of natural number, *as we now understand it*, that induction with respect to any well-defined property is a ground for asserting all natural numbers to have that property" (*ibid.*, p. 337; my emphasis). And Shaughan Lavine:[7]

> Part of what it is to define a property of natural numbers is to be willing to extend mathematical induction to it. To fail to do so is to violate our rules for extending and further specifying our arithmetical usage (*ibid.*, p. 231, n. 24).

Thought experiment: suppose that Karl accepts the axioms and rules of first-order arithmetic, say our theory A; suppose also that he is given a predicate P of the natural numbers. He does not know whether P can be defined in the language of A, but he does admit that P has a determinate extension over the natural numbers (or whatever the numerals denote). Now suppose that he refuses to apply induction to P or to formulas containing P until he is assured that P is definable in A. It is natural to conclude that Karl does not understand the natural numbers. What other conclusion is possible? Our truth predicate T is a case in point. Surely, the deflationist must agree that the extension of T is well-defined, since the extension is determined by the disquotation scheme or the T-sentences (provided only that the extension is limited to sentences or to natural numbers via Gödel numbering). That is all we need to extend induction. If someone understands arithmetic and admits that the extension of T (when restricted to $\mathscr{L}$) is well-defined, then he must admit formulas containing T into the induction scheme.

---

[6] "Reply to Wright," in *The Philosophy of Michael Dummett*, B. McGuinness and G. Oliveri, eds. (Dordrecht: Kluwer, 1994), pp. 329-38.

[7] *Understanding the Infinite* (Cambridge: Harvard, 1994).

Thus, in adding a truth predicate to our above theory A (to reach A^T), one is *not* really changing the subject or adding anything that was not there before. The idea is that the notion of arithmetic truth is somehow implicit in the arithmetic concepts. This much is congenial with the main, but vague deflationist theme that the truth predicate T refers to an insubstantial property. The problem is that, if our consequence relation is first-order, then we lose conservativeness in this natural move to A^T. The situation is reflected in a key notion in Dummett's writings, that of indefinite extensibility. Given any recursively enumerable set S such that every member of S is a truth of arithmetic, then one can effectively find a truth G_S of arithmetic that is not in the set. One way to do this is to add a truth predicate to the "theory" S and formalize the conclusion that every member of S is a truth of arithmetic. Dummett is correct that arithmetic truth is indefinitely extensible, at least in the technical sense used here. The question at hand is how this can be squared with deflationism.

There are as many—or more—versions of deflationism as there are deflationists. I cannot do an exhaustive survey, indicating how each one would feel the tension here or resolve it. We shall look at some of them.

I PRAGMATIC TRUTH

Let us start with an intriguing proposal made by some pragmatists. One criticism often leveled against deflationism is that it cannot distinguish areas of discourse that are objective and descriptive from those which are not, such as talk of taste or what is funny, or moral discourse on noncognitivist construals, or talk of causation on a Humean construal. The problem is that the disquotation scheme seems to apply to every declarative sentence, no matter what the subject is. 'Spinach is lousy', 'Cleese is funny', 'the cat is on the mat', and '7+5=12' are all true, or not, in the same, albeit thin, sense. Construed this way, the disquotation scheme is "promiscuous." Deflationists like Paul Horwich and Hartry Field[8] accept this promiscuity, and they each go to some length to soften the blow.

In a much discussed article, entitled "The Status of Content," Paul Boghossian[9] uses the promiscuity to argue that irrealism about semantic content is incoherent. He argues that, if there is a distinction between those areas of discourse which are objective and those which are not, then discourse about semantic content must be objective. Crispin Wright and Robert Kraut[10] independently respond that the Boghossian argument presupposes that the distinction between those discourses which are objective and those which are not is itself an objective distinction. That is, the Boghossian argument assumes that for any given discourse *D*, the claim that *D* is descriptive and objective is itself an objective, descriptive claim. There is a fact of the matter as to the objectivity of physics, mathematics, moral discourse, humor, and so on. Call the presupposition the *objectivity of objectivity claims* thesis. Kraut explicitly rejects the objectivity of objectivity claims thesis, proposing a pragmatist view in its place. According to Kraut, to call a statement 'true' is to call it objective, but that is only to manifest a certain stance toward the statement, the stance of taking-as-objective. This places Kraut with Robert Brandom, William James, and perhaps P. F. Strawson.[11] The idea is that the truth predicate does not function in a descriptive role. As Brandom puts it,

> The essential point to a theory such as James's is to treat calling something true as doing something more like praising it than like describing it.... Truth is treated, not as a property independent of our attitudes, to which they must eventually answer, but rather as a creature of taking-true and treating-as-true.... James's intent was...to mark off 'true,' like 'good,' as a term whose use involves the taking up of a nondescriptive stance, the undertaking of a commitment... (*op. cit.*, pp. 76-78).

Kraut tentatively suggests that the nondescriptive stance associated with truth may be something like "taking to be explanatorily ineliminable."

---

[8] Horwich, *Truth* (Cambridge: Blackwell, 1990), Field, "Disquotational Truth and Factually Defective Discourse," *Philosophical Review*, CIII (1994): 405-52.

[9] *Philosophical Review*, XCIX (1990): 157-84.

[10] Wright, *Truth and Objectivity* (Cambridge: Harvard, 1992), pp. 231-36; Kraut, "Robust Deflationism," *Philosophical Review*, CII (1993): 247-63.

[11] Brandom, "Pragmatism, Phenomenalism, and Truth Talk," *Midwest Studies in Philosophy*, XII (1987): 75-93; James, *Pragmatism: A New Name for Some Old Ways of Thinking* (Cambridge: Harvard, 1978); Strawson, "Truth," *Analysis*, IX (1948): 83-97, and "Truth," *Proceedings of the Aristotelian Society, Supplementary Volume 24* (1950): 129-56.

He shows how truth, so construed, need not be promiscuous. We do not manifest the stance of "treating as true" or "taking to be explanatorily ineliminable" to every asserted declarative sentence. A commonsense noncognitivist about humor, for example, will not take the relevant stance toward 'Cleese is funny', no matter how she is inclined toward Cleese's humor. Nevertheless, Kraut's view is still deflationary since the truth predicate does not express an objective property. He calls his view *robust deflationism*. It is a selective, prudish deflationism.

The present point about conservativeness is sharp when it is aimed at these pragmatic views on truth. Let us assume that arithmetic is objective. According to the pragmatist, the use of a truth predicate is to manifest a certain commitment, or stance, toward our favored arithmetic statements. We add such a predicate in order to provide the resources to express this commitment, or stance. But we just saw that adding a proper truth predicate allows us to prove theorems about the natural numbers—theorems we could not prove in the theory we had before we added the truth predicate. These new theorems only contain arithmetic terminology—they do not mention 'truth'. Thus, the new theorems are objective. From the pragmatist orientation, then, by adding resources whose only purpose is to manifest a certain stance, we are able to learn more about the objective realm of natural numbers. How can manifesting a stance (or stances) yield new objective knowledge?

One way out, perhaps, is for our pragmatist to deny that arithmetic is objective. This is not exactly a brand new philosophy of arithmetic, invented just to make the world safe for pragmatist deflationism. On such a view, presumably, our pragmatist would not add a truth predicate to arithmetic, since the only purpose of the truth predicate is to manifest the objectivity stance. Clearly, however, the truth predicate in this case is insightful. We learn things with it. Moreover, the consideration about conservativeness is quite general. Our pragmatist cannot accord objectivity to *any* theory that is as rich as elementary arithmetic. For example, if something resembling arithmetic can be simulated in a physical theory, then adding truth will not yield a conservative extension of the physics.[12] A pragmatist who takes this route must maintain that no rich area of discourse is objective—a hard pill to swallow and keep down.

II. TRUTH AND EXPLANATION

Other deflationists do not adopt the pragmatist line. They admit that truth is a predicate with an extension, but they hold that the property of truth is metaphysically thin, having no nature.[13] These deflationists must square their views with the fact that truth is not conservative over rich theories. As far as I can tell, the only retort for the deflationist (short of surrender) is to insist on a separation between semantic/metaphysical matters and epistemic/proof-theoretic matters. The deflationist must maintain that since truth is metaphysically thin, we cannot prove anything by adding a truth predicate that was not already implicit in the original subject. We may need a truth predicate (or similar device) in order to *see that* the Gödel sentence *G* or *Con* is a consequence of our original conception of the natural numbers, but *G* and *Con* are consequences of our original conception nonetheless.

---

[12] See my "Conservativeness and Incompleteness," this JOURNAL, LXXX, 9 (September 1983): 521-31

To maintain this line, the deflationist must employ a notion of consequence—or a notion of what is "implicit in" a certain conception—that is richer than what is yielded by first-order logic. I submit that deflationism requires a strong, inherently incomplete logic. I shall turn to some examples in the next section. Notice, however, that at least some deflationists are not in position to ignore the epistemic or proof-theoretic aspects of truth attributions. A common antideflationist argument turns on the explanatory success of science.[14] The idea is that the notion of truth plays a central role in a (or the) explanation of why science succeeds. Why is it, for example, that two successful scientific theories can often be combined into a more powerful, unified theory? The supposed explanation is that the successful theories are (approximately) true. The deflationist Michael Williams[15] concedes that reasons "for supposing that truth can or should play" an explanatory role "take us back in the direction of epistemology" (*ibid.*, p. 227). In present terms, the concession is that, *if* truth functions essentially in explanation, then the deflationist cannot remain in the realm of semantic consequence. He must maintain that truth is *proof-theoretically* conservative. The standard deflationist response is to claim that truth does not function essentially in explanation. Williams contends that the antideflationist argument succeeds "only if truth-talk has explanatory force and is not just a matter of expressive convenience. [The argument] will be persuasive only if we have already established an explanatory role for the concept of truth" (*ibid.*, p. 227). According to Williams, anything we invoke truth to explain, we can explain without it.

---

[13] The central claim of Grover, Camp, and Belnap's "Prosentential Theory of Truth" (*op. cit.*) is that truth is not (or need not be) a predicate. The issue is orthogonal to the theme of this paper, since they employ a strong logic in their prosententialist reconstruction.

[14] See Field, "Tarski's Theory of Truth," this JOURNAL, LXIX, 13 (July 13, 1972): 347-75 (which was written before Field became a leading deflationist); and R. Boyd, "The Current Status of Scientific Realism," *Erkenntnis*, XIX (1983): 45-90.

[15] "Do We (Epistemologists) Need a Theory of Truth?" *Philosophical Topics*, XIV (1986): 223-42.

Similar considerations apply to Kraut's pragmatic orientation to truth (*op. cit.*). According to his (tentative) proposal, to call a discourse objective—and to invoke a truth predicate for it—is to manifest the stance that statements in the discourse are explanatorily ineliminable. A crucial plank in Kraut's perspective is that assertions as to whether a discourse is objective—explanatory ineliminable—are not themselves objective claims. It follows, then, that Kraut himself manifests the stance that claims of objectivity play no explanatory role—they are explanatorily eliminable. Since a truth predicate is just a mark of an objectivity claim, Kraut is committed to the deflationist thesis that anything explained via the notion of truth can be explained without it.[16]

Go back to our theory of arithmetic *A* and its Gödel sentence *G* (or *Con*). Suppose that a logic teacher asserts that *G* is true, and a puzzled student asks for an explanation. The student accepts the teacher's word that *G* is true, but she wants to be shown why it is true. The student wants something like a convincing proof or an explanatory proof. The natural reply is to point out that all of the axioms of *A* are true and the rules of inference preserve truth. Thus, every theorem of *A* is true. It follows that '0=1' is not a theorem and so *A* is consistent. The Gödel sentence is equivalent to the consistency of *A*. It seems to me that this informal version of the derivation of *Con* and *G* is as good an *explanation* as there is. The argument shows why *G* is true. It does not help for the deflationist to point out that *Con* and *G* are (after all) semantic or logical consequences of the original theory *A* because this was the fact that needed explanation. Our student wanted to know why *G* is true—or why *G* is a consequence—and the move through the notion of truth provides the explanation.

Let us be clear about the uses of truth in this scenario. At the outset, the logic teacher asserts that *G* is true and the student wants to know why it is true. These uses of truth are indirect endorsements and are inessential. The teacher could (in some sense of 'could') assert *G* by writing 'I hereby assert...', followed by an instance of *G*. The student's query could also be made without invoking truth. The student could write 'Why is it that...' followed by an instance of *G* and a question mark. The crux of the matter lies in the use of truth in the proposed explanation. The teacher shows that every theorem of *A* is true. This use of truth—to make the generalization—is the centerpiece of the explanation, and the foregoing discussion focuses on it.

---

[16] Here I am indebted to Neil Tennant and Michael Dunn. Kraut himself uses the objective/nonobjective distinction (and thus the truth concept) in his global explanation of human verbal behavior.

In the scenario, we see a prima facie explanatory role for the notion of truth, if not in physics (as Boyd and the 1972 Field time-slice suggest) then in mathematics itself. Some deflationists might respond that this particular explanatory use of truth (in the generalization) does not undermine the claim that truth is metaphysically insubstantial. One might argue, for example, that only the use of truth in some sort of causal or empirical context is relevant to deflationism. After all, the target of the Boyd/Field argument is not mathematics, but rather the success of science at predicting and controlling the world. According to the retort under study, the deflationist backs off the bold claim that anything we invoke truth to explain we can explain without invoking truth. The retrenched claim would be that we do not need truth to explain anything empirical or causal.

In Lakatosian terms, this maneuver bars the monster. As invoked here, satisfaction (and thus truth) has a rigorous, formal characterization and presumably the mathematical context does not involve empirical or causal matters. The case concerning *G* does not advance the particular dialectic surrounding the Boyd and 1972 Field discussion. But the burden on a deflationist who takes this route is to delimit the selective attitude toward explanatory roles and to show why some such roles indicate semantic/metaphysical substance and some do not.[17]

A more bold, and straightforward response would be for the deflationist to maintain the general claim about explanation, and follow the lead of Williams and show how to get a decent explanation of *G* without a detour through truth. I suppose that the deflationist can embed the arithmetic *A* in a richer mathematical theory, such as set theory, and derive G in the richer theory. The claim would be that set theory provides the (real?) explanation, not arithmetic truth. Notice, however, that this "explanation" of G invokes resources far richer than arithmetic truth. The deflationist has invoked a ton of new ontology just to avoid the notion of truth in the explanation of G. Moreover, the usual proof of G in set theory proceeds by *defining arithmetic truth* in the richer theory (for example, set theory). Thus, this desperate maneuver does not really avoid the notion of truth. Furthermore, now that the deflationist has invoked set theory, we can construct a Gödel sentence G' for that theory and challenge the deflationist to explain G'. There is no avoiding the indefinite extensibility of truth for mathematical theories.[18]

---

[17] The issues advanced in Field's "Tarski's Theory of Truth" concern scientific method and reductionism According to the physicalist, any empirical/causal phenomenon must have a physicalistic explanation. The physicalist must show either that truth plays no explanatory role in empirical/causal contexts or that truth has a physicalistic reduction. The deflationist opts for the former. Since the Gödel sentence for arithmetic is prima facie not causal, it falls outside this particular debate. Notice, however, that our student very well might ask for an explanation of a Gödel sentence for a rich physical theory (see my "Conservativeness and Incompleteness"). I leave it to the principals to determine whether this context counts as empirical/causal.

There is, of course, no consensus on the notion of explanation, and so I do not claim that the present consideration is a knock-down argument against the deflationist claims about explanation. The philosophical literature contains many competing, and conflicting, accounts of scientific explanation. Some of these readily extend to mathematics, but others seem to preclude explanation in mathematics. For example, some accounts require a causal relation between *explanans* and *explanandum*. On an intuitive level, however, I submit that we do have a good explanation of G, and that this explanation invokes truth in the explanation. The burden is on the deflationist to show what is wrong with this picture.

III. BEYOND FIRST-ORDER CONSEQUENCE

Let us return to the notion of consequence needed for deflationism. Recall that the plan is to sustain the intuition that the Gödel sentence G is implicit in the original informal notion of natural number even though it cannot be derived using the original inferential resources, the theory A. First-order logic will not do, since, thanks to completeness, derivability and logical consequence coincide. If B is first-order, then every logical consequence of B can be derived from

---

[18] It is hardly worth mentioning that one cannot explain G by citing a theory which contains G as an axiom. This begs the question if anything does. Suppose that someone tries to explain G by citing a theory A^* consisting of A together with each instance of the reflection scheme. $\exists y \mathrm{PRF}(n,y) \rightarrow \Phi_n$. The sentence *Con* is an instance of this scheme and G is equivalent to *Con*. This "explanation" also begs the question. Why accept the reflection scheme? A separate justification for each instance is out of the question, since there are infinitely many such instances. Instead, we accept the reflection scheme when we realize that every theorem of A is *true*. That is, we notice that each instance of the reflection scheme is a theorem of A^T—each instance follows from (*) and a T-sentence. The consistency of A^* is provable in A^T. There is an affinity here with Anil Gupta's "A Critique of Deflationism," *Philosophical Topics*, XXI (1993): 57-81. Gupta inquires after the epistemic and semantic status of the T-scheme itself. We cannot learn or justify each instance separately.

B in a standard deductive system, without invoking new resources. I present three richer consequence relations here, as a representative sample: second-order logical consequence, conceptual consequence, and substitutional quantification.[19]

(1) Second-order consequence. The first option is for the deflationist to adopt second-order logic. He would replace the first-order induction scheme (in A) with the induction axiom, a single second-order sentence:

$$\forall X[(X0 \mathbin{\&} \forall x(Xx \rightarrow Xsx)) \rightarrow \forall x Xx]$$

The deductive system for the logic contains a comprehension scheme:

$$\exists X \forall x(Xx \equiv \Phi(x))$$

one instance for each formula Φ not containing X free. In the second-order framework, when new terminology is added to the language, formulas with the new terminology automatically go into the comprehension scheme, and the extensions of these formulas become instances of the induction axiom. Thus, *Con* and G are derivable, without further ado, as soon as the framework for truth is added to the theory.[20] There is no analogue of the above intermediate theory A'.

This approach presupposes something close to standard semantics for second-order logic, in which the predicate variables range over the entire powerset of the domain. This guarantees that the newly definable subsets of the natural numbers are within the range of induction. Second-order arithmetic registers the Dummettian idea that arithmetic truth is indefinitely extensible.

We saw that arithmetic truth is not conservative over arithmetic in the proof-theoretic sense. With second-order logic and standard semantics, however, we can articulate and sustain a model-theoretic version of the deflationist claim that truth is thin. Arithmetic truth is *semantically* conservative over arithmetic, in the sense that any model of second-order arithmetic can be extended to a model of arithmetic-plus-arithmetic-truth (in exactly one way).

---

[19] As far as I know, these three virtually cover the deflationist literature, and the literature on truth generally. Grover, Camp, and Belnap (*op. cit.*) and F. P. Ramsey invoke sentential quantification, which is a fragment of second-order quantification and similar to substitutional quantification; see Ramsey, "Facts and Propositions," *Proceedings of the Aristotelian Society, Supplementary Volume* 7 (1927), reprinted in Ramsey, *Foundations: Essays in Philosophy, Logic, Mathematics, and Economics*, D. H. Mellor, ed. (New York: Routledge, 1978), pp. 40-57; and "The Nature of Truth," *Episteme*, XVI (1991): 6-16. Another common option is to use an infinitary logic. Both sentential quantification and infinitary logic allow the expression of generalizations while avoiding truth. The invoked notions are rich and inherently incomplete, and so they do not avoid the considerations brought below (see my "Systems between First-order and Second-order Logics," *Handbook of Philosophical Logic*, Second Edition, forthcoming).

[20] See my *Foundations without Foundationalism: A Case for Second-order Logic* (New York: Oxford, 1991), chapter 5.

The result is general. Let Γ be any theory that can express its own syntax. Add a new predicate T to the language and add to Γ one of the common theories whose consequences are the T-sentences. Call the new theory Γ'. Then any model of Γ can be extended to a model of Γ'. Again, the T-sentences determine the relevant extension of T among the sentences of the original language.[21] It follows that Γ' is a semantic conservative extension of Γ: for any sentence Φ in the language of Γ (that is, any sentence that does not have any occurrences of T), if Φ holds in all models of Γ', then Φ holds in all models of Γ. In symbols:

$$\text{if } \Gamma' \models \Phi \text{ then } \Gamma \models \Phi$$

So second-order logic may be the salvation for the deflationist—if he can accept second-order consequence. Recall, however, that this consequence relation is very rich, much richer than arithmetic truth. There are second-order categorical characterizations of just about every major mathematical structure, including the natural numbers, the real numbers, the first inaccessible rank of the set-theoretic hierarchy, and beyond, well into the hierarchy of large cardinals. Therefore, truth for each of those theories can be reduced to second-order logical consequence. Moreover, there is a second-order sentence that is a logical truth if and only if the continuum hypothesis holds and another second-order sentence that is a logical truth only if the continuum hypothesis fails. Most opponents and most advocates of second-order logic agree that its consequence relation is mathematically deep and intractable. That battle concerns the significance of this observation. Among opponents, W. V. O. Quine[22] argues that second-order logic is not logic, but is "set theory in disguise," a wolf in sheep's clothing:

---

[21] There may be a complication due to the "extension" of T among nonsentences. One way to clean up this inelegance is to let sent(x) be a formula in the language of Γ that expresses 'x is a (name or Gödel number of a) sentence of Γ", and then add another axiom saying that only sentences are true: $\forall x(T(x) \rightarrow \text{sent}(x))$. In this scenario, however, the language cannot be first-order. It is a corollary of compactness that, if a first-order theory has infinitely many sentences, then there is no first-order formula whose extension (in all models) is isomorphic to the (names or Gödel numbers of) the sentences of the language.

[22] *Philosophy of Logic* (Englewood Cliffs, NJ: Prentice-Hall, 1986, second edition). See also I. Jané, "A Critical Appraisal of Second-order Logic," *History and Philosophy of Logic*, XIV (1993): 67-86, and S. Wagner, "The Rationalist Conception of Logic," *Notre Dame Journal of Formal Logic*, XXVIII (1987): 3-35.

> Set theory's staggering existential assumptions are cunningly hidden...in the tacit shift from schematic predicate letter to quantifiable variable (*ibid.*, p. 68).

Among advocates of second-order logic, Alonzo Church[23] writes that

> It is true that the...notion of [second-order] consequence...presupposes a certain absolute notion of ALL propositional functions.... But this is presupposed also in classical mathematics, especially classical analysis.... [L]ogic and mathematics should be characterized, not as different subjects, but as elementary and advanced parts of the same subject (*ibid.*, pp. 326n. and 332).

The point here is that no matter what one's views are on the status of second-order logic, a critic of deflationism might respond to the second-order maneuver by arguing that the deflationist is hiding the robustness of truth in the second-order consequence relation. If the consequence relation is itself robust, then the contemplated maneuver fails to show that truth is thin and has no nature. Once again, I do not know how the metaphysical robust/thin distinction is supposed to line up with the semantic, model-theoretic richness of second-order logic and we are in danger of losing our bearings.

(2) *Analytic consequence.* Let us move to our second option for the deflationist. Recall that we want the sentences *Con* and *G* to be consequences of our original theory *A* of arithmetic, even if they cannot be deduced therein. The puzzle is to delimit a plausible, deflationist friendly, notion of consequence that delivers this result. Our deflationist might invoke a notion of *conceptual consequence* or *analytic consequence.* This is the sense in which 'my apple is colored' follows from 'my apple is red'. The idea here is that *Con* and *G* somehow follow from the conception of the natural numbers or from the meaning of the arithmetic terminology. Once one understands what it is to be a natural number, one sees that the induction principle holds, and this understanding alone entails *Con* and *G*. We may need to invoke a notion like truth in order to see that *Con* and *G* are true, but they do follow from our understanding of the natural numbers alone. The information expressed by *Con* and *G* is implicit in the original conception of the natural numbers. Presumably, a deflationist who takes this option is unmoved by the Quinean attack on the analytic/synthetic distinction.

---

[23] *Introduction to Mathematical Logic* (Princeton: University Press, 1956). See also chapter 5 of my *Foundations without Foundationalism.*

There is a plausible intuition that there is no effective procedure to determine what follows from the meaning of a given proposition, but we can be much sharper here. I submit that every truth of arithmetic is a conceptual consequence of our understanding of the natural numbers. Suppose, for example, that Φ is a truth of arithmetic. What would it be for Φ *not* to be a conceptual consequence of our understanding of the natural numbers? Something like this: as far as our concept of natural number goes, it is possible for Φ to be false, or $\neg\Phi$ is compatible with our concept of natural numbers. But this is absurd. Our understanding of the natural numbers includes the full-induction principle, and that (with other principles) characterizes the structure of the numbers up to isomorphism. This structure, in turn, determines every arithmetic truth. Since by hypothesis, our proposition Φ is true, it holds in every such structure. In the jargon of possible worlds, Φ holds in all worlds that satisfy our conception of the natural numbers. It follows that conceptual consequence is inherently incomplete, and indefinitely extensible in Dummett's sense: given any effective procedure P that generates consequences of our conception of the natural numbers, we can effectively construct a consequence that is not generated by P.

Similarly, every truth of real analysis follows from our conception of the real numbers (which includes the completeness principle), and many of the truths of set theory, such as the size of the continuum, follow from our conception of the iterative hierarchy. Conceptual consequence may include second-order logical consequence, and it includes inferences that are not formal consequences (such as the above red-apple example). Thus, if our deflationist escapes from the foregoing dilemma by taking on board this notion of conceptual consequence, she has adopted a very complex and rich notion indeed. The question here is whether this notion is also metaphysically robust. I shall leave this for deflationists to resolve, if they are tempted by this route.

The reader may have noticed that I skirted the question of what the bearers of truth are. What (if anything) is truth a property of, or what (if anything) is 'is true' a predicate of? The plausibility of a rich notion of conceptual consequence turns on this question, because conceptual consequence arguably depends on semantic *content*. Conceptual, or analytic consequence is not a matter of form alone.

Some deflationists, like Ramsey and Horwich, hold that truth is a property of *propositions*. This view might support a notion of content yielding the requisite notion of conceptual consequence. It depends on what they go on to say about propositions, and whether propositions so construed have consequences sufficient for present purposes. Does the proposition that this apple is red entail the proposition that this apple is colored? Here, of course, the crucial question is whether the Gödel sentence follows from the content of the Peano postulates. If a deflationist adopts this rich notion of propositional content, a critic might retort that deflationism about truth is now uninteresting. All of the dreaded metaphysical substance is shifted elsewhere—to the notion of content. Even for these deflationists, there is a notion of *truth for sentences*, or truth for sentences in context: sentence *S* (in a given context) is true if the proposition expressed by *S* (in that context) is true. This *sentential* notion of truth is substantial, since it has a substantial component—the reference to propositions. It is small comfort that the other component—truth for propositions—is metaphysically thin and insubstantial.[24]

Other deflationists eschew propositions. Some take truth to be a property of sentence *types*, or perhaps sentence types paired with contexts. The types here are combinations of shapes and sizes; no semantic content is smuggled in. The most radical deflationist view eschews even types and holds that truth is a property of sentence *tokens* or *utterances*.[25] The radical view is suggested by an unrelenting nominalistic position that everything is physical. Accordingly, if truth is a property, it must be a physical property, and the best candidate for the range of this property is the realm of physical inscriptions and utterances.

I suggest that there is no hope for a notion of conceptual consequence on either of the views that reject propositions. First, what can it be for one *token* to be a consequence of another? In what sense can the physical hunks of ink or the sound waves that correspond to 'my apple is colored' follow from the physical hunks of ink or the sound waves that correspond to 'my apple is red'? If the waves on the beach happen to leave a token of the sentence 'my apple is red' in the sand, will it 'follow' that 'my apple is colored'? If they leave a token of the Peano postulates, will the Gödel sentence follow?[26] The move to types does not provide relief. In what sense can the shapes and sizes of the characters of 'my apple is red' produce or yield the shapes and sizes that correspond to 'my apple is colored'? How can the shapes and sizes of the characters of the Peano postulates yield the shapes and sizes of the characters of the Gödel sentence? If there is a coherent notion of conceptual consequence, it must hinge on semantic matters—on the fact that the Peano postulates *refer* to the natural numbers or that the axioms are categorical and semantically complete. The option of embracing conceptual consequence is not available unless our deflationist embraces something like semantic content.

---

[24] Here I am indebted to Wright and Sven Rosenkranz.

[25] See Field, "The Deflationary Conception of Truth," in *Fact, Science and Morality*, G. MacDonald and Wright, ed. (New York: Blackwell, 1986), pp. 55-117; "Disquotational Truth and Factually Defective Discourse," *Philosophical Review*, CIII (1994): 405-52; and "Deflationist Views of Meaning and Content," *Mind*, CIII (1994): 249-85.

[26] The radical nominalist/deflationist may also have a hard time with second-order logical consequence. How can this relation be understood as a physical relation connecting physical objects? For that matter, how is *deductive* consequence to be understood? I presume that anyone—nominalist, deflationist, or otherwise—needs some notion of consequence, even if they manage to get around the problem with Gödel's theorem.

In response, our deflationist might accept semantic content, but argue that it is thin and insubstantial, having no nature—along the lines of what he says about truth. To adapt Horwich's theme, he might claim that there is no more to the notion of content than is given by the infinitude of instances of the disquotation scheme:

'*P*' says that *P*

perhaps with some remarks about translation thrown in.[27] What directly matters here, however, is not the nature of propositions or concepts, but what *follows from* sentences, propositions, predicates, or concepts. Knowing that 'my apple is red' says that my apple is red and that 'my apple is colored' says that my apple is colored does not help in determining how or why the latter follows from the former. Adding some facts about how those sentences or the Peano postulates are to be translated into some other language does not help. Here again I lose my bearings, and so I leave this as another challenge to the deflationist. How can we have a rich notion of conceptual consequence without invoking a substantial notion of content?

(3) Substitutional consequence. The third, and final, notion of consequence to be considered here is associated with substitutional quantification. This one is especially pertinent since many deflationists employ substitutional quantification anyway, or something similar, such as sentential quantification or infinitary conjunctions and disjunctions. To see why, recall that a central use of a truth predicate is to express *generalizations* like 'everything the Pope says is true' or 'all of the theorems of *A* are true'.

Let us briefly review how generalization is accomplished on the standard, Tarskian development of truth. The key notion is *satisfaction*, a relation between (sequences of) objects and predicates or formulas with free variables. Satisfaction is a formal converse to the notion 'true of'. When we say, for example, that my apple satisfies the formula 'x is red', we are saying that 'red' is true of my apple. Tarski's framework is designed for so-called *formal* languages. The recursive definition of satisfaction follows the recursive characterization of the syntax. Let Sats(m,n) be a formula stating that the sequence m satisfies the formula coded by n. Let m be a sequence, x a variable, and o an object in the domain. Then let $m[x \mid o]$ be the sequence that assigns o to x and agrees with m on the other variables. For the purposes of generalization, the crucial clause in the recursive characterization of satisfaction is:

---

[27] See Field, "The Deflationary Conception of Truth."

$$\mathrm{Sats}(m, \ulcorner \forall x \Phi \urcorner) \equiv \forall o \mathrm{Sats}(m[x \mid o], \ulcorner \Phi \urcorner)$$

Suppose that Ψ has only x free. Then $\forall x \Psi$ is true if and only if Ψ is true of every object. Nothing could be more straightforward.

Notice that satisfaction does not add any techniques for generalizing per se. Instead, satisfaction allows us to employ a technique already in place—ordinary quantification. Anyone should regard this as a virtue.

Once again, if we expand the induction scheme to include formulas with the new Sats predicate (in the theory A''), then we can establish the generalization (*) that all the theorems of A are true. This is *Con*, which is equivalent to the Gödel sentence G for A. The Tarskian procedure is effective in the sense that, if the theorems of A are recursively enumerable, then so are the theorems of A''. In particular, A'' has its own Gödel sentence, which is neither provable nor refutable in A''. In other words, although A'' captures much of the notion of arithmetic truth, it does not allow us to establish of every truth that it is true. Thus, adding Tarskian truth to a theory does increase the resources, but the gain is modest. This is another virtue of the Tarskian framework.

It is interesting that there is no consensus among deflationists as to what Tarski did or showed, or whether he accomplished anything significant at all. Some say that Tarski's theory is of the dreaded, inflationary-correspondence variety. Horwich, for example, cites the "Tarski-inspired compositional approach" as a "substantial account of truth that is sometimes regarded as an elaboration of the 'correspondence theory'" (*op. cit.*, p. 117). He argues that in this Tarskian theory, the "extreme conceptual purity and simplicity of [deflationism] is thrown away for nothing" (*op. cit.*, p. 119). Tarski's theory is "unnecessarily complex, explanatorily misguided, and [fosters] the misleading impression that truth, reference, and satisfaction are inextricably linked with one another" (*op. cit.*, p. 120). See also Strawson (*op. cit.*). On the other hand, Field writes that

> Tarski's first contribution was to show how purely disquotational truth...is definable in...first-order logic, provided that sufficient ontology is available. Suppose that L is a fragment of our own language.... Then the obvious...Tarskian definition of truth-in-L, in our own language, is essentially the same as a purely disquotational notion of truth... (*ibid.*, p. 64).

Tarski himself was not interested in a philosophical analysis of truth, and I doubt if he cared whether truth has a substantial nature or not. He was also not out to capture everything there is to say about the use of the ordinary word 'true', since that use is probably inconsistent anyway. Tarski provided a formal characterization that aligns somewhat with ordinary usage, and which is consistent. As noted just above, his theory is designed only for formal, extensional languages that contain a bit of arithmetic and cannot express semantic notions. It is not clear how Tarskian satisfaction can be extended to more general languages, in order to increase the alignment with the ordinary usage of the truth predicate. Tarski's approach may not work in general, and so the deflationist recommends the raw disquotation scheme in its place. Without satisfaction, however, he will need another technique to express generalizations.

The most common device employed by deflationists is *substitutional quantification*. The Greek letter 'Π' is the usual symbol for a universal substitutional quantifier, and 'Σ' is the symbol for the corresponding existential quantifier. Let $\Phi(s)$ be a formula obtained from a complex sentence by removing one or more subsentences and replacing them with a (substitutional) variable s. Then $\Pi s\Phi(s)$ is true if for every sentence Ψ of the language, the result of substituting Ψ for s is true. Some deflationists, like Field, point out that this use of substitutional quantification is a device for infinitary conjunction. The sentence $\Pi s\Phi(s)$ amounts to the infinitary conjunction of all instances of $\Phi(s)$. This is the key to the logical behavior of these quantifiers.

The *disquotation scheme* is a single sentence of the substitutional language:

(DS) Πs('s' is true if and only if s)

or, if you like, an infinite conjunction of all of the T-sentences in disquotational form. This looks innocent enough, but looks can be deceiving.

As an aside, there are two issues concerning the range of the substitutional variable s. First, the strict deflationist, like Field, restricts application of truth to presently understood sentences (or perhaps sentences that can be translated into the present language). This is of a piece with the Quinean theme that truth is an "immanent" property, only applying to the current language. The restriction is counterintuitive. Suppose Karl says: 'Everything the guru says, or will say, is true':

Πs(If the guru ever says that s, then 's' is true)

This proclamation of faith is not supposed to be restricted to sentences that Karl (or the guru) presently understand. Suppose that at the time, neither of them knows a language as rich as that of set theory, but when they later learn some set theory the guru asserts the *continuum hypothesis* (CH). Karl's original proclamation applies, and so he would then believe CH (or else realize the error in his original proclamation). Thus, CH is included in the range of the original substitution variable s.

The second issue is whether the range of the substitution variable includes sentences containing the truth predicate (and the substitution quantifiers, and so on). If so, then there is a danger of self-reference and contradiction. Most deflationists recognize this, and agree that the substitutional scheme must be restricted in some way—but details are usually forgone. It is wise to forgo details, since there is no effective way to restrict the scheme and leave it sufficiently rich.[28] For present purposes, we avoid both issues and artificially restrict the substitution variable to the truth-free fragment of the original language.

Returning to the matter at hand, how do we get the relevant generalizations off the ground? Recall our deductive system A for arithmetic, and the crucial statement that all of the theorems of A are true. Recall the way we put this above, with the sentence:

(*) $\forall x(\exists y PRF(x,y) \rightarrow T(x))$

---

[28] See Vann McGee, "Maximal Consistent Sets of Instances of Tarski's Schema (T)," *Journal of Philosophical Logic*, XXI (1992): 235-41; and Alan Weir "Ultramaximalist Minimalism!," *Analysis*, LVI (1996): 10-22. If the language contains a truth predicate, then *every* sentence is provably equivalent to a T-sentence. The argument is a variation on the Curry paradox. Let Φ be a sentence and let Ψ be a fixed point for $Tx \equiv \Phi$, so that $\Psi \equiv (T\ulcorner\Psi\urcorner \equiv \Phi)$ is provable. But this last is equivalent to $(\Psi \equiv T\ulcorner\Psi\urcorner) \equiv \Phi$. So Φ is provably equivalent to the T-sentence $\Psi \equiv T\ulcorner\Psi\urcorner$. So the deflationist must omit at least one T-sentence corresponding to each false sentence of the language. How is he to know which sentences are false?

On the present account, this comes to:

(**) $\Pi s(\exists y PRF(\ulcorner s \urcorner, y) \rightarrow$ '*s*' is true)

with the substitutional variable *s* ranging over the sentences of *A*. The sentence (**) is sufficient for present purposes. If it could be established, then we would proceed as above. We substitute the sentence '0=1' for the variable '*s*' in the disquotation scheme (DS), concluding that '0=1' is not true. We then substitute '0=1' in (**), to derive *Con* and the Gödel sentence *G*.

Let us revive our subject Karl and ask how he can establish (**). According to the foregoing substitutional interpretation, (**) is just an infinite conjunction of all of its instances:

> If $\exists y PRF(\ulcorner 0=0 \urcorner, y)$ then '0=0' is true, and if $\exists y PRF(\ulcorner 0=1 \urcorner, y)$ then '0=1' is true, and if $\exists y PRF(\ulcorner$the successor function is one-to-one$\urcorner, y)$ then 'the successor is one-to-one' is true, and if $\exists y PRF(\ulcorner$the successor function is not one-to-one$\urcorner, y)$ then 'the successor is not one-to-one' is true, and....

Surely, Karl cannot establish (**) to his satisfaction by examining each of the conjuncts, since there are infinitely many of them. Karl cannot see, all at once, that every instance of (**) is true unless he already sees that *A* is consistent. Recall that we are not so much concerned with the truth of (**), but with how Karl comes to see that this generalization is true.

We saw above that the usual way to see that *A* is consistent is to note that the theorems of *A* are true. The axioms are either finite in number or else they are instances of a finite number of axiom schemes. In the latter case, it is reasonable to allow Karl to reflect on the schemes themselves, and see that all of the axioms are true:

> Πs(if '*s*' is an axiom, then '*s*' is true)

The same goes for the rules of inference. There are only finitely many rules of inference, and Karl can see that they all preserve truth:

> $\Pi s \Pi t \Pi u$(if '*s*' and '*t*' are true, and '*u*' follows from '*s*' and '*t*' by a rule of inference, then '*u*' is true)

As above, Karl remembers that the theorems of *A* are the closure of the axioms of *A* under the rules of inference, and so he justly concludes that all of the theorems of *A* are true:

> Πs(if '*s*' is a theorem of *A*, then '*s*' is true)

This is an informal or semiformal gloss on (**). Having got this far, Karl knows that A is consistent, in an informal or intuitive sense, and he can thus conclude that '$0=1$' is not a theorem of A also in an informal sense. He is not home yet, however. The informal statement, that '$0=1$' is not a theorem of A, is not in the language $\mathscr{L}$ of A (nor in $\mathscr{L}'$ for that matter). The sentence *Con*—$\neg\exists y\mathrm{PRF}(\ulcorner 0=1\urcorner, y)$—which is in $\mathscr{L}$, does not follow from the informal gloss on consistency or the informal gloss on (**). To establish *Con*, Karl needs (**) itself:

$\Pi s(\exists y\mathrm{PRF}(\ulcorner s\urcorner, y) \rightarrow$'s' is true)

and he does not have this yet.

Suppose that we expand the induction principle to include formulas with the substitutional quantifier, the truth predicate, the corner quotes, and even the quotation operator. Karl still cannot derive *Con* because the expanded theory does not sanction induction over *sentences* or numerals, only induction over natural numbers.

The crux of the matter lies with the antecedent of (**) and its informal gloss. We must pay attention to the delicate distinction between '"s" is a theorem of A' and $\exists y\mathrm{PRF}(\ulcorner s\urcorner, y)$. This is similar to the difference between the informal statement of consistency:

'$0=1$' is not a theorem of A

and *Con*, which is equivalent to $\forall y\neg\mathrm{PRF}(\ulcorner 0=1\urcorner, y)$. Derivations are certain finite sequence of sentences, and so fall squarely in the substitutional framework. The informal assertion of consistency comes to:

Πp('p' is not a proof of '$0=1$')

In contrast, $\forall y\neg\mathrm{PRF}(\ulcorner 0=1\urcorner, y)$ asserts that no natural number is the code of a proof of '$0=1$'.

The problem here is not with PRF versus "theorem of A." The formula PRF accurately renders the informal property of being a derivation in A. Again, $\mathrm{PRF}(m,n)$ is a theorem of A if and only if n is the Gödel number of a derivation in A of the sentence with Gödel number m. Moreover, $\neg\mathrm{PRF}(m,n)$ is a theorem of A if and only if n is not the Gödel number of a derivation in A of the sentence with Gödel number m. The problem also does not lie in the distinction between numbers and numerals. There is an intimate structural connection between strings of characters and natural numbers.[29] The Gödel theorem brilliantly exploits the connection, via Gödel numbering. There is an even tighter structural connection between numbers and numerals. The informal statement of consistency becomes:

---

[29] See John Corcoran, W. Frank, and M. Maloney, "String Theory," *Journal of Symbolic Logic*, XXXIX (1974): 625-37.

(*SubCon*) $\Pi n \neg PRF(\ulcorner 0=1 \urcorner, n)$

where the 'n' is a substitutional variable ranging over numerals. We now see that the contrast with *Con*:

$\forall y \neg PRF(\ulcorner 0=1 \urcorner, y)$

lies with the opening quantifier. There is a world of difference between the substitutional 'Π' and its objectual cousin '$\forall$'. In effect, *SubCon* amounts to:

> 0 is not the code of a proof of '0=1' and 1 is not the code of a proof of '0=1' and 2 is not the code of a proof of '0=1',....

In general, a universal quantification entails the conjunction of all of its instances, but not conversely. The infinite conjunction, or *SubCon* itself, does not logically entail *Con* because there is nothing in the infinite conjunction that entails that the denotata of '0', '1', '2', and so on exhaust the *natural numbers*. Indeed, there are models of *A* that contain elements not denoted by any numeral. Such models are called *nonstandard* and first-order logic does not have the resources to rule them out. The semantics for the substitutional quantifier rules out these models by fiat (like the standard semantics of second-order logic does).

If the deflationist is going to invoke substitutional quantification over sentences, instead of ordinary quantification over Gödel numbers of sentences (as in Tarskian satisfaction), then he has no reason to balk at the introduction of substitutional quantification over numerals and to connect that with ordinary quantification over numbers. That is, our deflationist should accept the obvious "bridge principle":

(*SubNum*) $\Pi n \Phi(n) \equiv \forall x \Phi(x)$

With this principle on board, Karl is home. *Con* and thus *G* are consequences of the expanded theory, with (*SubNum*) as an axiom. Induction over the natural numbers is the same as induction over the numerals and over the strings on the language—exactly as intended.

I apologize for the details, but they are needed to assess the consequence relation invoked by the deflationist. When he introduced substitutional quantification, he committed himself to the thesis that the range of the substitutional variables over sentences exhausts the "domain" of sentences. Similarly, substitutional quantification over numerals commits the deflationist to the thesis that '0', '1',...exhaust the "domain" of numerals. These are direct counterparts of the thesis that the denotations of '0', '1',...exhaust the natural numbers, and this last goes well beyond the resources of first-order logic.

Unlike the situation with Tarskian satisfaction via first-order logic, there is no Gödel sentence for Karl's expanded system with substitutional quantification. The reason is that the substitutional system (even without the principle *SubNum*) is not effective—not by a long shot. The consequence relation underlying substitutional quantification is much richer than that of first-order logic. The substitutional quantifier—like infinitary conjunction and disjunction—is a powerful logical device.[30] To see this, consider the so-called ω-rule for an ordinary system of arithmetic. It is an infinitary rule of inference:

from $\Phi(0), \Phi(1), \Phi(2)$,...infer $\forall x\Phi(x)$

A straightforward induction on the complexity of formulas shows that with the ω-rule (and the other axioms of arithmetic), one can "derive" all and only the truths of arithmetic. Thus, the consequences of such a system are not recursively enumerable. From Tarski's theorem the consequences of the ω-rule are not definable in arithmetic.

The semantic rule of a substitutional quantifier sanctions an ω-rule of sorts. It allows us to "infer" a (substitutional) universal generalization from its instances. If the substitution class is infinite, as with sentences or numerals, then the consequence relation is just as complex and intractable as the ω-rule. If the substitution is over numerals, then the relevant semantic rule is:

from $\Phi(0), \Phi(1), \Phi(2)$,...infer $\Pi n\Phi(n)$

This has the form of the ω-rule, and it is every bit as strong. If the above principle *SubNum* is added, then the ω-rule itself follows and our deflationist has invoked devices that yield, as consequences, every truth of arithmetic. It does not help for the deflationist to demur from *SubNum*. Suppose that Ψ is any sentence in the first-order language of arithmetic. Let Ψ' be the result of replacing each bound variable in Ψ with a substitutional variable over numerals, replacing each universal quantifier '$\forall$' with 'Π', and replacing each existential quantifier '∃' with 'Σ'. Then Ψ is true of the natural numbers if and only if Ψ′ is a consequence of the deflationist's theory. Thus, even without *SubNum* the deflationist has taken on an extremely complex and noneffective consequence relation. The resources of arithmetic are not sufficient to characterize this consequence relation. He has managed to "deflate" arithmetic truth only by invoking a quantifier whose consequence relation is at least as rich as arithmetic truth.

---

[30] For an early elaboration of this, see Dunn, and Belnap, "The Substitution Interpretation of the Quantifiers," *Noûs*, II (1968): 177-85. See also §9.1.4 of my *Foundations without Foundationalism* or §8 of my "Systems between First-order and Second-order Logics."

Recall that the Tarskian is not in this predicament. Tarski showed us how to add expressive resources to an ordinary, effective first-order theory of arithmetic—or any other extensional theory. We get the T-sentences as theorems, and perhaps the new theory is not conservative, but it is clearly not the case that for every arithmetic truth Ψ, a sentence equivalent to Ψ is a consequence of the expanded theory. To avoid whatever philosophical pitfalls the Tarskian may have, the deflationist has taken on a horribly complex, computationally intractable consequence relation.

I suppose that a deflationist might argue that substitutional quantification is needed anyway—independent of the issue of deflationism about truth. After all, Tarski's approach is limited and so perhaps even the antideflationist needs a rich device for generalizing about truth. I end at this juncture, with a challenge to the deflationist to make this case, and to show that substitutional quantification is metaphysically thin, despite being equivalent to arithmetic truth. There is an extra burden on the extreme deflationist who is also a physicalist/nominalist. This philosopher must show how we can make sense of a consequence relation which is every bit as complex as arithmetic truth and yet does not invoke anything abstract or otherwise nonphysical.

STEWART SHAPIRO

Ohio State University/Newark

PROPER NAMES, NAMES, AND FICTIVE OBJECTS

According to direct-reference theorists, proper names are used to pick out or mention an object without the intermediation of a tertium quid. There are two main variants of the theory, due respectively to Ruth Barcan Marcus[1] and Saul Kripke.[2] According to Marcus, proper names are tags; and according to Kripke, they are rigid designators. Although some commentators have treated these two notions as identical, Marcus has distinguished between them. As she has written: "Proper names are not assimilated to what later came to be called 'rigid designators' by Saul Kripke, although they share some features..." (*op. cit.*, p. xiii). She has not indicated why she discriminates tagging from rigid designation, but one can suggest an explanation. On her view, tagging is a property only of proper names and not of the names of natural kinds, whereas Kripke, in contrast, says that both kinds of names are rigid. This difference in the two versions may arise from the following conceptual source. Marcus holds that proper names are meaningless—"a proper name (of a thing) has no meaning" (*ibid.*)—and Kripke seems to agree—"There are writers, I think, who explicitly deny that names have meaning at all, even more strongly than I would..." (*op. cit.*, p. 32). But unlike Kripke, Marcus may also hold, or at least not rule out, that natural-kind terms are meaningful; and if that were the case, it would explain her refusal to identify tagging with rigid designation. One could thus express her position as maintaining that tagging is a narrower property than rigidity.

There is another difference in the two approaches. Marcus never gives a definition of a tag, or defines the nature of tagging. She does speak, however, expansively of the role tags play in language:

> If one views proper names as meaning-free, as tags for pure reference, then they too serve a quasi-indexical role, but with an advantage. On a causal or historical-chain theory of direct reference, proper names have an ostensive origin, are theoretically traceable causally or historically, and are part of a publicly available vocabulary of names with fixed reference in the language. Proper names serve as a long finger of ostension over time and place (*op. cit.*, p. 203).

---

[1] *Modalities* (New York: Oxford, 1993), p. xiii.

[2] *Naming and Necessity* (Cambridge: Harvard, 1981).

0022-362X/98/9510/522-34

Kripke, in contrast, provides an explicit definition of "rigid designator." It is a term that "in every possible world...designates the same object" (*op. cit.*, p. 48).

Despite these divergencies, however, the two theories have much in common. Both agree that by a proper name they mean what in ordinary language would be called a proper name. As Kripke says: "By a name here I will mean a proper name, i.e., the name of a person, a city, a country, etc.... We will use the term 'name' so that it does not include definite descriptions...but only those things which in ordinary language would be called 'proper names'" (*op. cit.*, p. 24). Their views thus differ sharply from that of Gottlob Frege, who uses 'name' in a broad sense to include declarative sentences, definite descriptions, common nouns, and indexicals, as well as proper nouns. Moreover, unlike Bertrand Russell,[3] who tries in a single sentence to define a proper name, that is, as "a simple symbol, directly designating an individual which is its meaning, and having this meaning in its own right, independently of the meanings of all other words" (*ibid.*, p. 174), they are both sensitive to the subtle forms that names take in ordinary speech. Russell's definition has all sorts of counterinstances: for example, that 'Bertrand Russell' is not a proper name because it contains two names, 'Bertrand' and 'Russell', and hence violates the criterion of being a simple symbol. But Marcus refuses to accept such a limited characterization. She points out that:

> In fact it often happens, in a growing, changing language, that a descriptive phrase comes to be used as a proper name—an identifying tag—and the descriptive meaning is lost or ignored. Sometimes we use certain devices, such as capitalization with or without dropping of the definite article, to indicate the change in use. 'The evening star' becomes 'Evening star,' 'the morning star' becomes 'Morning Star,' and they may come to be used as names for the same thing. Singular descriptions such as 'the little corporal,' 'the Prince of Denmark,' 'the sage of Concord,' or 'the great dissenter,' are, as we know, often used as alternative proper names of Napoleon, Hamlet, Emerson, and Oliver Wendell Holmes (*op. cit.*, p. 11).

Kripke makes similar comments:

> It should not be thought that every phrase of the form 'the x such that Fx' is always used in English as a description rather than a name. I guess everyone has heard about The Holy Roman Empire, which was

---

[3] *Introduction to Mathematical Philosophy* (London: Allen and Unwin, 1919), p. 174

> neither holy, Roman nor an empire. Today we have The United Nations. Here it would seem that since these things can be so-called, even though they are not Holy Roman United Nations, these phrases should be regarded not as definite descriptions, but as names (*op. cit.*, p. 26).

There are other similarities in their views about proper names. Marcus introduced the term 'tag' in arguing that a true identity sentence whose flanking expressions are proper names (tags) is necessary and not contingent. Kripke agrees with this claim. Both, for example, reject W. V. Quine's idea that, if one tags the same object twice, and by different names—say the planet Venus by 'Phosphorus' and 'Hesperus'—the identity sentence, 'Phosphorus is Hesperus', is contingent. According to Kripke, Quine was confusing the concept of contingency with that of a posteriority. It was an a posteriori discovery that one had tagged Venus twice, but since necessarily Venus is Venus, it follows that 'Phosphorus is Hesperus' is necessary and not contingent.

Although the terminology of 'tagging' is not found in the early literature on proper names—say, in such writers as John Stuart Mill, Frege, and Russell—the notion that a proper name is applied to an individual and sticks to that individual in all circumstances, possible and actual, has a venerable history. In *Philosophical Investigations*, Ludwig Wittgenstein[4] says: "To repeat—naming is something like attaching a label to a thing" (*ibid.*, §26).

No doubt, Wittgenstein inherited this notion from Russell; and indeed as both Marcus and Kripke have acknowledged, Russell is the source of many of their ideas about reference. Of course, they have modified Russell's views, rejecting the idea that proper names are meaningful and also the idea that proper names are abbreviations for definite descriptions. They also disagreed with Russell's contention that a sentence like 'Scott is the author of Waverley' is not an identity sentence. What they did accept from Russell is that proper names must be distinguished from descriptions—that these locutions play radically different logical roles; and accordingly, that one cannot substitute descriptions for names in every context in which the latter occur. As is well known, Russell eventually held that logic requires the existence of *genuine* proper names. These on his view turn out to be demonstratives and indexicals, such as 'this', 'that', 'here', and 'now'. Such genuine proper names function, in effect, like tags, and require ostension at some juncture. As Marcus sees it, Russell was saying, in effect, that a proper name requires a referent:

---

[4] G. E. M. Anscombe, trans. (Cambridge: Blackwell, 1958).

> On this account "proper name" is a semantical, not a merely syntactical notion. Reference is supposed. We may mistakenly believe of some syntactically proper name, say "Homer," that it has an actual singular referent and is a genuine proper name, but if its use does not finally link it to a singular object, it is not a genuine name at all (*op. cit.*, p. 203).

Thus Russell, Marcus, and Kripke all agree that a designator cannot be a (genuine) proper name, unless there exists a particular individual that it tags. As we shall see in a moment, the idea that there is a direct link between such a name and its referent has important ramifications for such theories in speaking about fictive objects. Marcus is very clear on this point. She states that:

> It is not, I will propose, the general absence of "identification *conditions*" that makes possibilia problematic. It is that possibilia cannot be objects of reference at all (*op. cit.*, p. 204).

Fictive objects are possibilia in Marcus's sense, so they cannot be the objects of reference.

In concluding this brief historical survey, I should mention that Kripke has at times used Marcus's term 'tag' as a synonym for 'proper name'. In a discussion reprinted in *Modalities*, he states "tags are the 'essential' denoting phrases for individuals, but empirical descriptions are not," and says similar things elsewhere (*op. cit.*, pp. 34-35). Accordingly, I shall assume in the remainder of this essay that their views about proper names are similar, namely, that such names are tags.

I. THE PROBLEM OF FICTION

The aforementioned theory is very powerful and has attracted a large number of adherents, among them such distinguished philosophers as Hilary Putnam, David Kaplan, Keith Donnellan, Zeno Vendler, and Howard Wettstein. Among its merits are: its defense of true identity sentences (flanked by tags) as necessary; its ability to distinguish clearly between such notions as analyticity, necessity, a priority, and a posteriority; its sharp contrast between names and descriptions; and its avoidance of talking about a possible world as if it were "a foreign country, a distant planet way out there," as Kripke puts it. Talk about possible worlds on Kripke's account is always translatable into ordinary discourse, and becomes talk about counterfactual situations involving tagged entities. Thus, his theory does not require a special interpretation of the role of proper names in counterfactual contexts.

Yet despite its many merits, the view has a special difficulty concerning the names of nonexistent entities. Kripke was much aware of the problem. In *Naming and Necessity*, he says:

> Concerning rigidity: In many places, both in this preface and in the text of this monograph, I deliberately ignore delicate questions arising from the possible nonexistence of an object (*op. cit.*, p. 21, footnote).

That "delicate" question is troubling to Donnellan,[5] too. In his interesting variant of the direct reference account—the so-called *historical-explanation theory*—he distinguishes between issues involving the nonexistent from those involving fictive objects. The latter, of course, are a subclass of the nonexistent. Yet he implies that there is no adequate theory involving the fictional nonexistent:

> What is to be excluded from consideration here is an account of discourse about fiction. (This is not, of course, to say that such an account is not in the end needed.) (*ibid.*, p. 218).

On all versions of the direct-reference theory, a serious problem about the status of fictive names arises. If fictive objects have what in ordinary discourse would be called proper names, such as 'Odysseus', 'Lancelot', and 'Shangri Lai', and if, as everyone would agree, such objects do not exist, then there is no way of tagging them. You cannot attach labels to the nonexistent, as the previous quotations from Marcus make plain. So, clearly, their names cannot be tags or rigid designators. The conceptual model with which direct-reference theorists are working presupposes that to tag or label something requires not only an act of the tagger (or speaker) but also the existence of the object to be tagged and labeled. One cannot put a label on a nonexistent piece of furniture. Given this outlook, the direct reference theorist thus faces a conundrum. These are among the available options for the theory:

(1) Abandon the thesis that proper names are tags. A suboption here might be to distinguish between proper names as tags and proper names as rigid designators. In effect, this would be to adopt Kripke's version of the theory and not Marcus's.

(2) Adopt a variant of Russell's theory of descriptions. Unlike the canonical version of the theory of descriptions that maintains that all ordinary names are abbreviations for descriptions, this variant would hold that in fiction, so-called proper names are abbreviations for descriptions, while maintaining that in ordinary, nonfictive discourse, they are tags. This option would have two consequences: it would not accept the theory of descriptions as applying to the names of existing objects; and, secondly, it would result in a nonho-

---

[5] "Speaking of Nothing," in S.P. Schwartz, ed., *Naming, Necessity, and Natural Kinds* (Ithaca: Cornell, 1977).

mogeneous theory about the role or roles of names in ordinary speech.

(3) Argue that in fiction, so-called proper names have a symbiotic, derivative, or secondary use. This option would differ from the previous as follows. To say that fictive names have a derivative use does not imply that they have to be analyzed as definite descriptions. Both Kripke (*op. cit.*, pp. 26-29) and Donnellan, for example, have explicitly denied that any proper names are abbreviations for descriptions. A fictive proper name would thus be what ordinary speech regards as a proper name, only its referential function would not be direct.

(4) Maintain that discourse about fictive objects presupposes a *fictive existence operator*, so that fictive objects do exist in fiction, and therefore can be tagged by fictive proper names. Similar remarks could be made about heraldry. Tagging would thus be relativized to a particular body of discourse. In espousing this position, one could appeal to a theory of contextual implication, differing versions of which are found in the work of H. P. Grice, John Searle, and A. P. Martinich.

This set of options does not exhaust the possibilities, of course, but it does bring out the nature of the dilemma that the direct-reference theorist faces; for none of these options is attractive. To abandon the notion that names are tags is, of course, simply to abandon the theory. To draw a distinction between tagging and rigidity for proper names would require something that at least at present does not exist: a new theory that would explain how rigidity differs from tagging.

To accept the Russellian account that fictive names are descriptions is to abandon the Kripkean notion that proper names are what ordinary persons describe them as being. No ordinary speaker would accept that 'the great detective described in the mystery stories written by A. Conan Doyle' is the name of Sherlock Holmes.

To adopt the notion that fictive names are derivative or symbiotic, and are about fictive objects in some sense of 'about' (such as the secondary sense of 'about' that Peter Strawson mentions) seems simply to beg the question. This also seems to be the difficulty with Vendler's[8] account.

To accept the fourth option that one can tag fictive objects in fiction obviously involves an extended or quixotic use of 'tag'. What seems to be undeniable is that one cannot tag the nonexistent.

Let us summarize the problem. It arises because the following theses, all of which are essential to the direct-reference theory, seem to be in conflict.

---

[8] *The Matter of Minds* (New York: Oxford, 1984), chapter 4.

(1) Proper names are tags (or rigid designators).
(2) Tags (or rigid designators) require the existence of something to be tagged.
(3) Fictive objects do not exist, and therefore cannot be tagged.
(4) Fictive objects in ordinary discourse have proper names.
(5) A proper name is what is ordinarily meant by a proper name.

II. THE ORDINARY PERSON'S PERSPECTIVE

I shall suggest a solution to the problem. In setting the stage, let us move away from theory for a moment and consider how ordinary persons think and talk about names. Of course, one must concede at the outset that such persons do not think very deeply, if at all, about the problems that vex philosophers of language. It is also dubious that such folk have anything that could be called a *theory* about proper names. Still, they do talk about, and on occasion reflect deeply, about the names they choose for all sorts of things: children, pets, ships, and airplanes. Moreover, the ordinary use of names is much more complex and subtle than most theorists have acknowledged. Take, as an example, the claim made by Marcus that "linguists exclude proper names from the lexicon altogether" (*op. cit.*, p. 203). This is a contention with which many philosophers agree, among them Kripke and Paul Ziff.[7] But are they right? Just turning to the Ms in my dictionary, I find entries for a large number of what they call proper names, including the names of fictive entities: Mastricht, Macao, Macaulay, Machiavelli, Madame Butterfly, Medusa, Madrid, and Malaga. Some philosophers, as we have seen, have held that proper names are meaningless. Yet, dictionaries tell us that 'Sarah' in Hebrew means "princess," that 'Svensen' in Norwegian means "the son of Sven," and that 'Momotaro' in Japanese means "the son of a peach." Some philosophers have even stated that proper names do not belong to any language. Yet, clearly 'Giovanni' is Italian, 'John' is English, and 'Ivan' is Russian (of course, the same name—for example, 'Paul'—may belong to more than one language). Indeed, ordinary speakers may well have important things to tell us about names that philosophers have overlooked or ignored. If one were to ask such a person what she thinks about the five theses that seem to be in conflict, one would elicit some instructive answers.

First, a reflective, thoughtful person would certainly reject step (5). This states that a proper name is what is ordinarily meant by a proper name. That person would point out that in ordinary speech, the term 'proper name' is hardly ever used. Clearly, the ordinary person would be right. 'Proper name' is a technical term, invented by grammarians, linguists, and other specialists in language. In such disciplines, it is used metalinguistically to speak about the grammar of particular object languages, and in particular to distinguish nouns that apply to individual persons, places, and things from so-called common nouns, like 'tiger' and 'gold'. Ordinary speakers are not in the grammar business and hence do not employ (or need) this metalinguistic terminology. Accordingly, they do not say of persons or places that they have proper names. Instead, they speak about the names simpliciter of such entities and objects. They will ask: 'What is the name of the city you visited last year?' not 'What is the proper name of the city you visited last year?'; or 'What is his name?' and not 'What is his proper name?' An ordinary person hearing the question 'What is his proper name?' would be puzzled by it. Is the interrogator suggesting that the person is using a pseudonym or an alias, for example? Apart from such special contexts, the question would possibly not be understood at all.

---

[7] *Semantic Analysis* (Ithaca: Cornell, 1967), pp. 86-89, 93.

Second, they would accept premise (4), that is, that fictive objects in ordinary discourse have proper names, only if it were modified. The modification would consist in dropping the word 'proper' and thus concurring that in ordinary discourse, fictive objects have names. That is, they would say that words like 'Sherlock Holmes' and 'Odysseus', are just as much names as 'Bill Clinton' or 'Bertrand Russell'. If asked by a philosopher of what 'Sherlock Holmes' and 'Odysseus' are supposedly the names (and assuming that they know the relevant literary references), they might be puzzled by the question. It seems obvious that they are the names of Sherlock Holmes and Odysseus.

Before considering the other three steps in the argument, one might also pursue the point further, in a Meno-like fashion, seeing if one could elicit an argument in support of step (4) as modified. Here are two that might emerge from the process.

(1) Suppose a philosopher were to ask an ordinary individual to pick out the proper names in the following list:

'Paula'
'the discoverer of America'
'a metal'
'Santa Claus'
'Columbus'
'Zeus'
'chief of staff'

Assuming that they understood the question, since it employs the locution 'proper name', such persons would select 'Paula', 'Santa Claus', 'Columbus', and 'Zeus' and would reject 'the discoverer of America', 'a metal', and 'chief of staff' as names. One cannot imagine naming a new born child 'the discoverer of America', for example. Since two of the selections in the list are the names of fictive entities, Santa Claus and Zeus, one can conclude that, for ordinary speakers, such entities have what philosophers would regard as proper names.

(2) Now a second argument. Suppose a person were looking through a book of names in order to select one for a newborn male child. That list might include 'Sherlock', 'Lancelot', and 'Hamlet'. These are names made famous in Western fiction; yet they can be used as the names of existing persons. So why think they are not names when used in fiction? A name is a name is a name in any context, Gertrude Stein might have said. An ordinary person would find no obstacle in identifying these words as names. Playing the philosophical game, we could even add that they are proper names.

Direct-reference theorists would probably not accept this argument. They might contend that it rests on a confusion, namely, that a list or a book of names does not contain proper names in a strict sense of the term; for those are the actual names of existing individuals, whereas what a book gives us are only potential names. Potential names become proper names only when they are applied to particular persons, and then they function as tags. It is therefore misleading to think of potential names as proper names per se. This response is interesting, but I believe it begs the question. It assumes that which is in dispute: that *N* is a proper name only if *N* is applied to an existing person, place, or thing. It is therefore not a compelling objection.

III. THE AXIOM OF REFERRING

We shall resume our inquiry by considering the remaining three premises of the argument from a philosophical perspective, and thus leave the ordinary man to go about his daily activities. I accept premise (3) that fictive objects do not exist, and therefore cannot be tagged. I also accept premise (2) that tags require the existence of something to be tagged. Marcus is clearly right in asserting both (3) and (2). Fictive objects are mere possibilia and possibilia cannot be tagged. I also accept the ordinary person's modifications of premises (5) and (4). So in order to provide a solution to the direct-reference's dilemma which arises from accepting all five premises, we must reject premise (1), namely, that proper names are tags. In explaining why, we must begin by reiterating what it is that direct-reference theorists mean in describing a proper name as a tag. A proper name is used—and here the terminology varies depending on the theory—for identifying, picking out, speaking about, mentioning, or referring to a particular person, place, city, object, and the like. Let us speak for the moment only about referring. The theory thus rests on a principle that is assumed by all direct-reference theorists and even by many non-direct-reference theorists, such as Strawson and Searle. I call this the *axiom of referring.* It can be formulated as follows:

> One cannot tag (that is, directly refer to) that which does not exist.

Although not a direct-reference theorist, Searle's[8] version is almost identical with the preceding. He writes:

> Whatever is referred to must exist (*ibid.*, p. 77).

There are a number of variants of the axiom. Strawson's[9] version states:

> ...it would not *in general* be correct to say that a statement was about Mr. X, or the-so-and-so, unless there were such a person or thing (*ibid.*, p. 35).

We find a still different, more radical formulation in a recent work by Newton Garver and Seung-Chong Lee.[10] They say:

> In philosophy, it is no good talking about something that is not there (*ibid.*, p. 115).

The principle can be traced back to the pre-Socratics, but its locus classicus in philosophy of language can be found in Frege. In a famous passage he asserts:

> He who does not acknowledge the nominatum cannot ascribe or deny a predicate to it (*ibid.*, p. 90).

I call this principle, with its variations, an *axiom* because it is never in fact supported by arguments, reasons, or evidence. It is simply assumed to be true. It is then imported into theories of reference and plays a crucial role in them. Yet, although accepted as true by virtually all philosophers of language, it is patently false. The irony of the situation is not merely that it is false, but that it is obviously so; for it is a plain fact that we do use language to refer to nonexistent (including fictive) objects by name, and to make true (or sometimes false) statements about such objects. It is plainly true to say that Odysseus was married to Penelope, and that Hamlet was not married to Ophelia, and false to say otherwise. Literary critics, and indeed just ordinary readers, frequently discuss the characters of fictive personages: for example, Emma in the eponymous novel by Jane Austen. In doing so, they use all the resources of everyday speech that they would use in talking about real persons, alluding to places where such personages live, identifying those places by name, historical origin, and so forth. Those fictive names are transferred by a causal chain from generation to generation in just the way that the names of real historical persons are.

---

[8] *Speech Acts* (New York: Cambridge, 1969).

[9] "On Referring," in A. Flew, ed., *Essays in Conceptual Analysis* (New York: Macmillan, 1960), p. 35.

[10] *Derrida and Wittgenstein* (Philadelphia: Temple, 1994).

In such contexts, one is speaking about fictive objects not in any secondary sense of 'about', if there be such, but in the same sense of about in which I have been speaking about Saul Kripke and Ruth Marcus in this paper. More generally, what is true is that whatever can be said about nonfictional objects can be said about fictional ones; and whatever can be said in our daily talk outside of fiction can be said within or about fiction. Such talk includes every possible use of ordinary discourse: jokes, lies, true and false statements, direct and indirect references, asides, and the application of names to characters, places, and things.

IV. THE SOLUTION

What is the significance of the preceding remarks for direct-reference theories? I submit that we should abandon the axiom of referring in any of its forms, since it is palpably false. This would clear the way for a better account of names, including what the tradition calls "proper names," and also of statement making. In its place, I shall now propose an alternative account that has manifold advantages over direct-reference theories and none of their liabilities. This will be my solution to the dilemma described earlier.

First, the account would entail that the names used to refer to fictive entities, such as Sherlock Holmes and Odysseus, really are names, and are not to be distinguished from the names of nonfictive entities, such as Ruth Marcus and Saul Kripke. One would thus have a uniform, rather than a fractionated, account of names. Second, it would entail that we should abandon the notion that proper names are tags. I have agreed that, if a tag is a label that one applies to an object, one cannot tag things that do not exist or that are mere possibilia, as Marcus correctly states. Thus, abandoning the notion that proper names are tags (premise 1 of the dilemma), we can then use such names to refer to fictive entities in just the way that ordinary speakers do. We would thus need no special theory about the use of so-called proper names in fiction.

Third, it would entail that we do not have to construe so-called proper names as abbreviated descriptions in fictive contexts. Even better, it would entail that such names are not abbreviations for descriptions whether used in fictive or nonfictive contexts. Whatever the context, they would just be names in the ordinary sense of that term. The Russellian and the direct-reference options would thus both be rendered otiose. Fourth, it would entail that there is no difference in referential power between names and descriptions. The idea that so-called proper names refer directly and that descriptions do not is a red herring. Misidentification can occur with the use of names and in other ways as well. I can refer with "equal directness" to a neighbor as 'the young man next door', or as 'Donald Wilson'. Which I do shall depend on various contextual factors, such as whether I know the name of the person next door or not. But naming has no special referential sanctity in such cases. Depending on circumstances, either can do the job equally well.

Fifth, it would entail a greater sensitivity to the nuances of everyday discourse. One of the serious liabilities of the direct-reference doctrine is that it attempts to find one and only one relationship—tagging—to explain how names refer. My view says that reference via names is multifarious, having various forms. In particular, we should distinguish between such notions as mentioning *x*, and referring to *x*; and between both of them and such notions as picking out *x* and identifying *x*, which should also be discriminated from one another. These are all different actions. I can pick out persons, say, males from females in a group, without being able to identify any particular individual, and each of these actions is different from mentioning or referring to someone or something. I can perform both of the latter without picking anything out or identifying anything or anyone. I can mention by name, say, that Jack the Ripper was the serial killer who murdered *X* without being able to identify Jack the Ripper, and certainly without being able to pick him out of a group of suspects in a police lineup. Picking out, mentioning, speaking about, referring to, identifying, and other modes of discriminating things from one another are all ways that names work in everyday speech. There is thus no single form of specification, such as tagging, which is the key to understanding how names are used in human communication.

Sixth, we can dismiss the Meinongian fantasy that fictive objects must exist in some sense or other or we could not name, refer to, or make true statements about them. My view has the advantage that fictive objects are acknowledged without hedging as nonexistent. Finally, we do not need any such artificial ways of categorizing language as to say that there are assumed "existence operators" for fiction, mythology, or heraldry. Supporters of such a program would ask us to say such things as: '"Odysseus" is a proper name in fiction', or 'That Odyssesus was married to Penelope is true' means that this statement is true in fiction. My account does not divide the language into such labyrinthian and artificial compartments. Ordinary language is one, indivisible body of discourse which we use to speak about many different kinds of things, including mythological, heraldic, and fictive entities. In using it for such purposes, we need no existence operators for marking off spurious boundaries.

I believe that the axiom of referring is the source of much philosophical pathology. It is interesting to speculate on why philosophers have almost universally and uncritically accepted it. No doubt, a full explanation would require another essay. But it is obvious that it is their ontological and metaphysical concerns that lead them to the mistaken ideas that proper names are tags and that only existent entities can have proper names. Their philosophy of language is thus driven by ontological considerations. If we begin instead from the perspective of language itself—asking what counts as a name in everyday speech and how such names are used—all sorts of philosophical misconceptions will vanish. It is this sanitized conception of philosophy of language which I have been urging in this essay.

AVRUM STROLL

University of California/San Diego

BOOK REVIEWS

A World of States of Affairs. D. M. ARMSTRONG. New York: Cambridge University Press, 1997. xiii + 285 p. Cloth $54.95.

In this wide-ranging book, D. M. Armstrong has woven together his various metaphysical doctrines into a systematic whole founded on the version of realism about properties for which he is justly famous. For him, properties are universals, not entities existing in some Platonic heaven, but contingent existents firmly rooted in the natural world, being wholly located wherever they are instantiated (and they do all have instances). His realism is doubled, since besides universals, he also admits states of affairs ('facts' in some mouths, where these are distinct from true propositions): if the particular *a* instantiates the universal *F*, there is a third entity, the state of affairs *a's being F*.

Armstrong's metaphysics is conducted within a naturalistic framework ("the world, the totality of entities, is nothing more than the spacetime system" (5)) which gives it the flavor of an open-ended research program. We would expect trouble trying to force the subject matters of mathematics, modality, and morality into the naturalistic mould. Armstrong offers us revisions of his previous accounts of the first two, but there is nothing on the third. Since he rejects regularity theories of singular causation and laws, in favor of something stronger, there are also three chapters explaining how a naturalistic account can be given of the very stuff of science. I shall concentrate on the ontological basics and, in passing, it will be noted how Armstrong's respect for the natural sciences, which underpins his naturalism, leads him to refashion metaphysics itself.

If it were not preposterous, it would be something of a scandal that Armstrong now advertises his conception of universals as *Fregean*:

> The universal is a gutted state of affairs; it is everything that is left in the state of affairs after the particular particulars involved in the state of affairs have been abstracted away in thought....
>
> This contention will at once recall Frege's doctrine of the unsaturatedness of his 'concepts'. I happily acknowledge the influence, and, indeed, think of his concepts as close relatives of my universals (28-29).

In one place, "Fregean" becomes "Fregean-Aristotelian" (202), an impossible metaphysical alliance, so we know something is up. Gottlob Frege's concepts, the referents of predicates, are abundant despite

0022-362X/98/9510/535–40

their extensional identity conditions. There are concepts that actually have instances, concepts that actually have none but might have had some, and even concepts that cannot have any. Thus the Fregean doctrine has nothing to do with Armstrong's claims that science discovers what universals there are and that every universal has an actual instance. Moreover, it is a mistake to impute an ontology of states of affairs to Frege. In the mature work in which the idea of the incompleteness of concepts appears, he can find no room in the realm of reference for states of affairs. Notoriously, in his semantics all true sentences name the same object—the true—and so the differentiation required by an ontology of states of affairs can only be found at the level of sense, namely, among the true propositions. No doubt there is much that is obscure in his projection of the incompleteness of predicates onto the concepts to which they refer. But this incompleteness has nothing to do with the idea that a universal is a "gutted state of affairs." Readers of Frank Ramsey's "Universals"[1] will know that the idea of universals as gutted states of affairs does nothing to distinguish them from particulars. For, adapting Armstrong's characterization of a universal (29), we may say that a particular is a state-of-affairs type, the constituent that is common to all states of affairs that contain that particular. Particulars are as unsaturated as universals in Armstrong's system, which is more evidence that it is not Fregean.

Armstrong also recommends that universals be thought of as ways things are, which is supposed to help us see that it is wrong to "*substantialize* universals" (30), to provide a motive for denying that universals are capable of independent existence ("How can ways exist by themselves...?" (99)) and to see that universals with no instances are peculiar ("For they will be ways that are ways nothing is. Can we take such ways with ontological seriousness?" (154)). Here, Armstrong's fondness for rhetorical questions is illustrated, combined with what seems like an ordinary-language argument appealing to our use of 'way'. But appearances are deceptive, since he has chosen to speak of ways so that he can get the right results. In response to Sydney Shoemaker, he admits that universals might be thought of as "ways things *can* be" (38), which characterization would allow independently existing and uninstantiated universals. So he rules it out on the ground that universals would then be necessary existents (not a transparent inference) and he is having none of those. Skeptics about the significance of the distinction between particular and universals will say that particulars are just as much ways universals are as universals are ways particulars are, to which Armstrong will reply, in propagandist fashion, that this does not accord with his use of 'way'.

---

[1] Reprinted in his *Philosophical Papers*, D. H. Mellor, ed. (New York: Cambridge, 1990), pp. 8–30.

According to Armstrong's use of 'universal', it is "a very simple but very illuminating analytic truth...that universals are strictly identical in their different instantiations" (27). But the evident power of the truth should make one suspicious of its alleged simplicity, for it is used to argue that a simple universal *could not* be complex (33) and to deliver the "Principle of Instantial Invariance" ("a particular universal can only have one -adicity" (85)). The latter principle is understood in a peculiar way, since it is held to rule out a binary relation which sometimes takes two different objects as its arguments and sometimes the same object twice over (208-09). How do we get this principle from the original analytic truth? By linguistic sleight of hand: "How could a three-term relation be strictly identical with a two-term relation?" (85). This does nothing to rule out so-called "multigrade universals," just as the question 'How can a universal located in one place be strictly identical with one in another?' does nothing to rule out universals being wholly present in more than one place.

Like Humpty-Dumpty, Armstrong is free to make 'universal' mean what he chooses and hence to create the analytic truths he needs to deliver the metaphysical principles he wants. The real question is whether what he says has any bearing on what exists. Distinguish three questions: Are there any universals? Which sorts of universals are there? Which universals of which sorts are there? We shall come to Armstrong's a priori answer to the first. His answer to the third has always been resolutely a posteriori: it is not for us armchair metaphysicians with our ordinary-language and a priori arguments to decide the exact composition of reality. No, that is a question for empirical science. This leaves the second question. Universals can be classified in various, orthogonal ways, one of which mirrors the complexity of predicates resulting from negation, disjunction, and conjunction. Twenty years ago, Armstrong[2] thought he had an a priori argument for rejecting the existence of disjunctive universals, such as *P or Q*, which used the analytic truth mentioned earlier:

> Suppose *a* has a property P but lacks Q, while *b* has Q but lacks P. It seems laughable to conclude from these premises that *a* and *b* are identical in some respect (*ibid.*, p. 20).

In this book, he tries the same argument but, realizing that it begs the question, falls back on the metaphysician's favorite, an appeal to economy ("it is hard to see that there *has* to be one" (27)). Then he retreats once again claiming that selected disjunctive universals might be accepted "*if the development of the natural sciences appeared to demand this*" (28). This change of mind marks a distinct shift in emphasis in his metaphysical method. In earlier work, his naturalism, combined with the insistence that science discovers the catalogue of universals, were enough to make metaphysics empirically respectable, with plenty of room left for the metaphysical imagination. But now a bland appeal to the unknown results of the sciences is used time and time again to block in the argument where the metaphysics runs out. Just when to make the appeal seems arbitrary. For example, returning to the issue of multigrade universals, Armstrong has said nothing to rule out a scientific decision on their existence.

---

[2] *Universals and Scientific Realism*, Volume II (New York: Cambridge, 1978).

As I understand him, Armstrong argues for the existence of universals, the answer to the first question, by using what he rates "perhaps the fundamental argument of this book" (115), the truth-maker argument. In the first place, this argument creates a theoretical role which is best realized by his states of affairs but, since universals are constituents of these states of affairs, it is also an argument for their existence. The argument uses a principle, every truth has a truth maker, which embodies a version of the correspondence theory of truth but puts a particular spin on the underlying commonsense realism, according to which we are not the masters, representing does not make it so, and hence if our thoughts and utterances are to be true, the world must be a certain way. The latter condition is turned into the demand that there must be some *entity* in the world which will "ensure, make true, underlie, serve as the ontological ground for" (116) any given truth. And the truth-making relation is spelled out in terms of a noncausal relation of necessitation:

> ...the truthmaker for a truth must necessitate that truth. ...if a certain truthmaker makes a certain truth true, then there is no alternative world where that truthmaker exists but the truth is a false proposition (115).

Not quite a definition of truth making, but nowhere is one explicitly provided. But everything Armstrong says suggests that the quoted passage states a necessary *and* sufficient condition: an entity makes a proposition true if and only if the existence of the entity entails that the proposition is true (I ignore Armstrong's naturalistic worries about propositions). States of affairs are needed as truth makers, since a sentence of the form 'a is F' can be false even if a and/or

Fness exist but not if the state of affairs *a's being F* exists. This does not mean that each truth has its own personal truth maker, since a posteriori realism about universals dictates the same for states of affairs: two or more sentences may have the same truth maker and one sentence may have more than one truth maker. Nor does it follow that truth makers are always states of affairs: for example, '*a* is identical to *a*' is made true by *a* alone.

There is much that ought to be said about truth making and the associated principle but, unfortunately, Armstrong does not say it. First, there are obvious gaps in the slide from simple realism to the heavy-duty metaphysics required by the truth-maker principle. For example, one might worry whether a way the world is need be an entity in the world and whether the explanation of truth making in terms of necessitation is richer and more controversial than anything supplied by common sense. Second, it is extraordinary that Armstrong does not consider an obvious version of the slingshot argument, which aims to show that every truth has the same truth maker (even more extraordinary when one notes that Armstrong has read a book on facts which devotes a chapter to slingshots[3]). Third, the application of the truth-maker principle to contingent *and* necessary truths needs justification. After all, the realist thought with which we started is only clearly applicable to contingent truths. We think: a contingent truth might have been false and so there must be some difference between the way things are and the way they might have been. But Armstrong does apply it across the board and runs into trouble. Anyone familiar with relevance in logic will spot trouble with his account of truth making. According to it, any entity will make every necessary truth true. So the search for the truth makers of mathematical truths rests on the misguided principle that "Different necessary truths have different truthmakers" (183). And his investigation of the *different* truth makers of modal truths is equally wrongheaded, since if 'Necessarily P' and 'Possibly P' are themselves necessarily true if true (and Armstrong hopes they are (170)), any entity will make every modal truth true.

I end with some remarks on style. In the Preface, Armstrong tells us that he is "no lover of footnotes" (xi) and hence none appears in the book. I share Armstrong's aversion to some extent, but not his therapy. Endless footnotes distract the reader, encourage hedging and pseudo-scholarship, and provide a dumping ground for scraps of research which one cannot bear to waste. It is their content, not just their presence, which is pernicious. But Armstrong is only concerned with appearances, merely incorporating in the main text what would have been consigned to the foot. He thus finds it necessary to apologize to those readers who "feel that certain passages should have been relegated in this way" (xi): a master stroke of perversity.

---

[3] Kenneth Olson, *An Essay on Facts* (Stanford: CSLI, 1987).

I cannot help regretting that Armstrong has abnegated the freshness and vitality of his earliest work in metaphysics. He now thinks we need to know precisely who said what to whom, who told him what others had written, and when they all said it ("as an undergraduate at Sydney University" (16)). I am inclined to think that this is not a mere obsession with the intimate details of his philosophical biography but a sign that he has assumed the persona of the grandee with all the patronage which that entails.

ALEX OLIVER

Queens' College/University of Cambridge

For the Sake of the Argument: Ramsey Test Conditionals, Inductive Inference, and Nonmonotonic Reasoning. ISAAC LEVI. New York: Cambridge University Press, 1996. xv + 341 p. Cloth $64.95.

In the first half of his latest book, Isaac Levi expounds a theory of the revision of beliefs and suppositions, and elaborates on it to take into account bodies of belief and supposition that include inductive inferences. Induction has not recently had many friends among philosophers, but it has found some among logicians and computer scientists who study so-called nonmonotonic reasoning. The second half of the book is mainly taken up by a critical discussion, from the decision-theoretic perspective of the earlier chapters, of the merits of some of their proposals.

The account of belief revision of Carlos Alchourrón, Peter Gärdenfors, and David Makinson serves Levi as a foil. Arguing that it is not really an account of belief revision at all, insofar as it does not take up the question of when it is rational to adopt a belief, he notes that this objection disappears if we take the object of study to be not belief change, but instead the making of suppositions. Supposing is like pretending, or making believe, in that suppositions do not call for justification in the way that beliefs do. We can make them for the sake of the argument.

Beliefs and suppositions are studied indirectly, as represented in *corpora*, or sets of sentences. One can be said to believe or suppose every-

0022-362X/98/9510/540–46

thing to which his beliefs and suppositions commit him; in keeping with this, Levi's corpora include all of their logical consequences. The making of suppositions for the sake of the argument is studied by considering revision operations that introduce suppositions into corpora.

Levi distinguishes two main operations. In *Alchourrón/Gärdenfors/Makinson revision*, to suppose A for the sake of the argument, you must first pretend not to believe $\neg A$ (if indeed you do believe $\neg A$); then you incorporate A. In *Ramsey revision*, you pretend not to have beliefs about A one way or the other, before incorporating A. Both kinds of revision factor the making of suppositions into two operations that Levi considers more basic: pretending not to believe something is *contraction*, the removal of a sentence from a corpus. The subsequent incorporation of a sentence is *expansion*.

Expansion is straightforward in Levi's view. To incorporate a sentence, one adds it to a corpus and closes the result under logical consequence. Contraction is another matter, since in general there is more than one way to remove a sentence. Levi argues that those potential contractions are to be preferred which minimize the value of lost information, and, as a first attempt at making this idea precise, he proposes to evaluate potential contractions using the probability-based measure of informational value which is familiar from his writings on induction (and is also discussed later in this book). This measure satisfies a requirement that Levi thinks it ought to, namely, that the value of a *stronger* (that is, more inclusive) corpus is to be at least as great as the value of a weaker one.

This first attempt is unsuccessful. As Levi explains, it turns out with this measure that all preferred contractions that remove A are *saturatable*, in the sense that subsequent incorporation of A will yield a corpus that is maximally consistent. Clearly, this will not do. With this measure, one who supposes something for the sake of the argument will end up committed one way or another on every question, since, for every sentence, either the sentence itself will be in his corpus or its negation will be. This difficulty leads Levi to propose, for the purpose of choosing between potential contractions, a different measure of "damped" informational value.

Any given potential contraction is the intersection of certain saturatable contractions. Its *damped* informational value is the least of the informational values (in the earlier sense) of these. Using this measure, Levi is able to recommend, among potential contractions, the weakest that minimizes loss of informational value (it is the intersection of all maximally valued saturatable contractions). This move to damped informational value does seem to avoid the earlier difficulty.

But given the central place of contraction in Levi's ideas about revision, the interpretation of conditional sentences, and more, one might well hope for more motivation than he offers.

Here is my attempt to provide it. Arguably, to believe fully is to take one's beliefs to be correct. Furthermore, one who takes his beliefs to be correct must expect that acquiring further correct information will lead him to *add* to them. Of course, this expectation will be disappointed if it turns out that some of his beliefs are false. Still, to believe fully is to expect that stronger corpora are the alternatives between which, if all is well, one will be led to choose. Now, where alternatives carry values of one kind or another, a choice between them can, in turn, be evaluated. Its value depends on the values of the alternatives between which it is a choice. In particular, we might suppose, the informational value of a corpus depends on the values of the corpora that are stronger than it is. Assuming in addition, as Levi does, that the value of a stronger corpus is at least as great as that of a weaker one, it follows that a corpus inherits (at most) the *least* of the values of the stronger corpora. Damped informational value can be regarded as value that is inherited by potential contractions in this way—from stronger saturatable contractions.

With this infrastructure in place, Levi turns in chapter 2 to the interpretation of conditional sentences. He follows Frank Ramsey's suggestion that to argue about a conditional is hypothetically to add its antecedent to one's stock of knowledge, and then to argue on that basis about the consequent. Along these lines, he proposes *Ramsey tests*: a conditional $A>C$ is *acceptable* in relation to a corpus of beliefs and suppositions if C is in the result of introducing A into the corpus; otherwise, $\neg(A>C)$ is acceptable. In this way, Levi obtains an acceptability condition for conditional sentences from his account of supposing for the sake of the argument.

Conditionals do not appear within Levi's corpora. He argues that while they can be acceptable or unacceptable, they do not carry truth values and therefore have no place within corpora representing states of full belief or supposition. So, whereas, say, 'This piece of plastic will melt' can be supposed for the sake of the argument, in his view (strictly speaking), 'This piece of plastic will melt, if heated' cannot be. I shall soon say more about Levi's reasons for thinking this. Meanwhile, notice that excluding conditionals from corpora greatly reduces the scope of the Ramsey test, which does not cover *iterated* examples, such as 'If this piece of plastic will melt if heated, then it is unsuitable as an insulator'. To assess the acceptability of a conditional using the Ramsey test, we must introduce its antecedent into a corpus; but this we cannot do in such a case. The point is related to another made by Peter Geach,[2] in connection with his criticism of Gilbert Ryle's view that a conditional is not of a kind with its consequent and antecedent, but is a license to infer the one from the other. Geach also sees a difficulty in the case of conditionals with complex antecedents.

Iterated conditionals seem perfectly intelligible, though, so Levi proposes to extend the scope of his acceptability condition. Although we cannot suppose a conditional to be true, we can for the sake of the argument suppose it to be acceptable. He develops this idea so that some iterated sentences gain acceptability conditions that are in the spirit of the Ramsey tests. The suggestion is fine as far as it goes, but it does not go far enough. Since Levi does not say what it is for disjunctions and other compounds of conditionals to be supposed acceptable, his suggestion does not cover more deeply embedded examples, such as 'If this piece of plastic will melt if heated, or will crack if cooled, then it is unsuitable as an insulator'. We might ask what can be done to extend further the scope of Levi's acceptability condition. Instead, we shall see why he thinks conditionals lack truth values, since this is what leads him to treat iterated ones differently in the first place.

Take an epistemic modal sentence, say, 'It is seriously possible that A' ('$\Diamond A$'). We can think of it as a conditional interpreted according to the Ramsey test, since, where T is any given truth of logic, $\neg(T>\neg A)$ is acceptable in relation to a corpus, under commonly made assumptions, precisely if A is a serious possibility in relation to it. Now, according to Levi, this sentence does not carry a truth value. As we shall see, he thinks its doing so is at odds with the reasonable notion that error is to be avoided in inquiry. He argues that change must introduce error, if epistemic modals carry truth values and appear within corpora.

Take, for example, change that incorporates new information. In relation to a corpus that does not include A, $\neg A$ is a serious possibility. So, as Levi sees it, if judgments of serious possibility carry truth values, then such a corpus will include $\Diamond\neg A$. Consider some given corpus that does not include A, but is compatible with A; and suppose we introduce A into it. The resulting corpus cannot include $\Diamond\neg A$, since, in relation to it, $\neg A$ is not a serious possibility. It seems that on incorporating A, $\Diamond\neg A$ must be replaced by $\neg\Diamond\neg A$. Now Levi writes:

---

[2] "Assertion," *Philosophical Review*, LXXIV, 4 (1965). 449-65

> If all the sentences in [the earlier corpus] are judged true by the inquirer, changing his theory must perforce deliberately introduce error.... If avoidance of error is a desideratum in inquiry, sentences of the form "It is seriously possible that h" and "It is not seriously possible that h" should not be considered to carry truth values... (55).

In short, Levi argues that to replace $\Diamond \neg A$ by $\neg \Diamond \neg A$ on incorporating A is deliberately to replace a sentence that one is certain is true by another that one is certain is false.

This argument is inconclusive. If we are going to consider the possibility that judgments of serious possibility carry truth values, then we must allow them to do so in relation to corpora. When evaluated in relation to the corpus that does not include A, $\Diamond \neg A$ will be true; but in relation to the corpus that does include A, instead $\neg \Diamond \neg A$ will be true. Now, if we allow this, the inquirer who replaces $\Diamond \neg A$ by $\neg \Diamond \neg A$ on incorporating A does not deliberately introduce error. He does not introduce error at all, because the corpus changes in relation to which these sentences are to be evaluated. What he does is to replace what *was* a truth (but has become a falsity) by what has *become* a truth (but was a falsity), and this is just what is needed to *avoid* error. Once the contextuality of epistemic modals is allowed for, it will be seen that Levi's argument does not establish that they lack truth values. Iterated and noniterated conditionals might after all be treated on a par.

In chapter 5, Levi turns to *inductively extended* corpora that include inductive inferences. Induction, broadly construed, has long interested researchers in artificial intelligence; they call it *nonmonotonic reasoning*. Levi takes on a proposal about nonmonotonic reasoning due to Makinson and Gärdenfors, according to which a conclusion C can be inferred from a premise A, relative to some corpus K, if C is included in the revision of K to incorporate A. (This is just the condition under which $A>C$ is acceptable in relation to K.) Levi argues at length that this relation between A and C cannot be regarded as inference. His point is an important one, and it has not been fully appreciated by students of nonmonotonic reasoning. It ties in with differences between inference relations and conditionals that have been noticed before. For example, whereas the acceptability of a conditional can vary from corpus to corpus, inference relations are not contingent. Whether a conclusion can be inferred from a premise ought not depend on one's beliefs, as Makinson and Gärdenfors allow it to do.

Later sections of chapter 5 explore the revision of inductively extended corpora, and its formal resemblance to revision according to Alchourrón, Gärdenfors, and Makinson. The discussion does not presuppose any particular account of induction but, in chapter 6, Levi soon turns to the decision-theoretic, parametric account which he prefers, and which is familiar from his earlier books. Later chapters take up a comparison of this account with proposals about nonmonotonic reasoning which have arisen in computer science, or which have been inspired by the developments there. One objective is to see whether these proposals can be, as Levi puts it, "rationalized" by the comparison—whether they can be shown, given suitable settings of parameters, to share salient properties with his own account of induction. I shall get to one of these properties soon.

Ray Reiter's *default logic*, which rests on the notion of a default rule of inference, is a first candidate for the comparison. Levi also considers a proposal of David Poole, which can be regarded as a restricted form of default logic. (John McCarthy's *circumscription* he leaves to one side.) Neither Reiter nor Poole introduces statistical considerations explicitly, although they seem to rely on them for informal motivation. Among proposals that explicitly introduce statistical methods, Levi argues with much vigor against the view that sentences are to be accepted whose credal probabilities are sufficiently high.

The sources of nonmonotonicity that have emerged within artificial intelligence research are diverse: they include competition between arguments, reasoning about prototypical cases, rules of inference by default, and more besides. Not surprisingly, a range of formally different proposals have been made about nonmonotonic reasoning. Perhaps embarrassed by these riches, Dov Gabbay, Makinson, and others have suggested formal properties they think sensible proposals should have, and have sometimes invoked them to recommend one proposal above another.

Levi discusses how the characteristics of his account of induction vary as parameter settings are changed. For example, with suitable settings, it fails to display the "cautious monotonicity" that has been thought proper: adding a conclusion drawn by induction to the premises from which it was drawn can lead to a change in the conclusions that can be drawn by induction. Levi somewhat gingerly suggests that, in the light of this, the failure of a proposal about nonmonotonic reasoning to secure cautious monotonicity might not appear to be a defect.

To my mind he is quite right. His discussion illustrates the wrong headedness of using these properties to adjudicate between proposals. To use them in this way puts the cart before the horse, since, although one usually can gain a good idea of whether any given proposal is sensible by thinking about the intuitive ideas that underlie it, one's sense of whether the relevant formal properties are desirable is, prior to having thought about particular proposals, quite poor. It seems to me that, if a proposal is firmly grounded in informal ideas but fails to display cautious monotonicity, as Levi's does, then so much the worse for cautious monotonicity.

Levi considers in his discussion a *corrected* default logic. He sees a need for a charitable interpretation, since, he claims, with the standard formulation the default extension of a set need not be closed under logical consequences, and logically equivalent sets need not have the same default extensions (208). In fact, the charity is unneeded, since both claims rest on misunderstandings. Levi also claims that default logic is not cautiously monotonic (209). Indeed, it is not, but the demonstration is incorrect (there is another, correct demonstration in note 3 on page 317; but mind the typographical error). I have traced the errors, which ramify throughout chapter 7, to the preliminary version of a secondary source upon which Levi unfortunately relies. At the relevant points, he cites the published (and corrected) version. While many technical details of this chapter need to be revised because of the errors, the main points do not. This is a case of charity returning home: without his charitable interpretation, Levi would have seriously misrepresented default logic; but as things stand, he does not. The reader may wish to have at hand Reiter's original discussion, as well as Makinson's survey.[3]

Logicians and computer scientists who work on nonmonotonic reasoning have been casual about what it is that justifies such reasoning. Apart from the occasional gesture toward a broadly pragmatic rationale, they have in the main left such questions to one side. By taking up the comparison with his own explicitly decision-theoretic proposal, Levi has placed questions of justification on the agenda, and indicated where answers might be sought. This is in itself valuable. Philosophers will find in this book a conscientious and by and large successful attempt to come to terms with some of the basic work on nonmonotonic reasoning, and to render it in language that is familiar to them.

MICHAEL MORREAU

University of Maryland/College Park

---

[3] Reiter, "A Logic for Default Reasoning," *Artificial Intelligence*, XIII (1980): 81–132; Makinson, "General Patterns of Nonmonotonic Reasoning," in Gabbay, C. J. Hogger, and J. A. Robinson, eds., *Handbook of Logic in Artificial Intelligence and Logic Programming, Volume III: Nonmonotonic Reasoning and Uncertain Reasoning* (New York: Oxford, 1994), pp. 35–110.

THE JOURNAL OF PHILOSOPHY
VOLUME XCV, NO. 11, NOVEMBER 1998

THE PROBLEM IN TWENTIETH-CENTURY PHILOSOPHY*

The title refers to the notorious problem Ludwig Wittgenstein[1] raises very near the end of the *Tractatus*:

> It is clear that the logical product of two elementary propositions can neither be a tautology nor a contradiction. The statement that a point in the visual field has two different colours at the same time is a contradiction (*ibid.*, section 6.3751).

These two statements, each plausible by itself, seem none too plausible when taken together:

(1) The spot is red and blue.

(1) is "the logical product of two elementary propositions," and hence, according to the first statement, (1) cannot be a contradiction; but it asserts that "a point in the visual field has two different colours at the same time," and hence, according to the second, (1) is a contradiction.

(1) together with sentences like (2)-(6) seem to present us with a conflict between intuition and logic:

(2) The spot is red.
(3) The spot is not blue.
(4) The spot has a color.
(5) Red is a color.
(6) The spot is green.

---

* Thanks to Arthur Collins, Hans G. Herzberger, Jaegwon Kim, David Pitt, Peter Ross, George E. Smith, Ernest Sosa, Virginia Valian, and Adam Vinueza

[1] *Tractatus-Logico-Philosophicus*, D. F. Pears and B. F. McGuiness, trans (New York: Routledge, 1922).

0022-362X/98/9511/547–75

Intuition tells us that (1) is a contradiction, that (2) and (6) are incompatible, that (2) entails (3) and (4), and that (5) is a necessary truth. But logic appears to contravene intuition. Not only do the sentential components of (1) express elementary propositions, (2), (4), (5), and (6) themselves express elementary propositions, and (3) expresses the negation of an elementary proposition. Since elementary propositions are logically independent of one another, logic tells us that sentences (1)-(6) do not have the logical properties and relations that intuition tells us they have.

Wittgenstein[2] imposed both a methodological and an epistemological constraint on a solution to this problem:

> It must be possible for the contradiction to show itself entirely in the symbolism, for if I say of a patch that it is red and green, it is certainly at most only one of these two, and the contradiction must be contained in the *sense* of the two propositions (*ibid.*, p. 107).

The methodological constraint is straightforwardly expressed in the first clause of this conjunction. The epistemological constraint, that knowledge of the contradiction is a priori semantic knowledge, is expressed in the second. Insofar as the symbolism shows the semantic properties and relations to be, as Wittgenstein says, "contained" in the senses of the sentences, it explains our knowledge of the contradiction, and of relations such as that (2) entails (3), as grounded on nothing more than an understanding of the sentences.

The problem (1) raises is widely known under names like the *color incompatibility problem*, but it is no more a problem specifically about color words than it is a problem specifically about the property of contradiction. It is a general problem about the extralogical vocabulary of the language and about all the semantic properties and relations of the language. The problem surfaces whenever we try to explain the logical powers of extralogical words with a symbolism on which the logical form of elementary propositions affords no basis for their explanation. But even though the color vocabulary is only a special case of this general problem, the conflict between our intuitions about color words (and other sensation terms) and the logic of sentences containing them is by far the hardest case, and for this reason, it serves as the test case for solutions to the general problem.

I. EXISTING SOLUTIONS TO THE PROBLEM

The definite article in the title is not intended to suggest that the work of twentieth-century philosophy is on the verge of completion. The uniqueness of Wittgenstein's problem lies rather in the special role it has played in shaping twentieth-century philosophy. The emergence of the problem was one of those events which totally transforms an intellectual discipline. The three movements in which most analytic philosophy of this century has been done, Wittgenstein's late philosophy, Rudolf Carnap's neo-Humean empiricism, and W. V. Quine's neo-Millian empiricism, were each, in large measure, responses to the problem. When we appreciate the extent to which these movements were shaped by attempts to solve or dissolve the problem, we shall appreciate how long a shadow it casts over twentieth-century philosophy.

---

[2] *Philosophical Remarks* (Chicago: University Press, 1975).

I.1. ***Wittgenstein.*** In the years following the completion of the *Tractatus*, the problem was the test case for its semantic system. Wittgenstein's failure to show a contradiction like (1) within that system and his subsequent diagnosis of the failure led to his radically different way of thinking about meaning, language, and logic in the late philosophy. To my knowledge, Edwin B. Allaire[3] was the first to recognize the central role of the problem in the development of Wittgenstein's late philosophy. Nowadays, it is widely recognized among Wittgenstein scholars. Referring to Wittgenstein's attempt in "Some Remarks on Logical Form"[4] to revise the truth-table notation to express "exclusion," Max Black[5] says: "Here, it may be said, Wittgenstein's system begins to crack" (*ibid.*, p. 368). P. M. S. Hacker and Anthony Kenny,[6] two other well-known Wittgenstein scholars, take essentially the same view of the centrality of the problem in the transition from the early to the late philosophy.

Wittgenstein's[7] opposition to synthetic a priori knowledge led him to equate 'being synthetic' with 'being contingent', and, as a consequence, this put him under great pressure to take sentences like (1) to be meaningless (*ibid.*, p. 68). But as he still held a calculus picture of language during this period, the option of representing sentences like (1) as meaningless, while at the same time representing their negations as meaningful truths, was not open to him.[8] Since the only kind of sentential well-formedness in calculi is syntactic well-formedness, the meaningful/meaningless distinction coincides with the well-formed/ill-formed distinction. Hence (1) and its negation are either both meaningful or else both meaningless.

---

[3] "'Tractatus' 6.3751," *Analysis*, XIX, 5 (1959): 100-05; reprinted in Irving Copi and Robert W. Beard, eds., *Essays On Wittgenstein's* Tractatus (New York: Macmillan, 1966), pp. 189-94.

[4] *Aristotelian Society, Supplementary Volume IX* (1929) 162-71, here p. 168, reprinted in Copi and Beard, pp. 31-38

[5] *A Companion to Wittgenstein's* 'Tractatus' (Ithaca. Cornell, 1964).

[6] Hacker, *Insight and Illusion* (New York: Oxford, 1972), pp 86-94; Kenny, *Wittgenstein* (Cambridge: Harvard, 1973), pp. 103-19

[7] *Ludwig Wittgenstein und der Wiener Kreis. Notes of F. Waismann and B. McGuinness* (Malden, MA: Blackwell, 1967).

[8] "Some Remarks on Logical Form," pp. 170-71.

Allaire suggests that the need to separate meaningfulness from syntactic well-formedness led Wittgenstein to abandon the calculus picture of language and develop his new use-based conception of meaning:

> One possible way of making it palatable [that (1) but not its negation is meaningless] is to identify meaning with use. This, as we all know, is one of the key ideas of the final phase. If I am right, it follows that with respect to one very major issue, at least, the 1929 paper marks a transitional stage between the thought of the *Tractatus* and that of the final stage (*op. cit.*, pp. 192-93).

Since, on the new conception, the meaningfulness of signs is an aspect of their use in our linguistic practices, it is possible to reject the identification of the meaningful with the syntactically well-formed. Hence the negation of (1) can be treated as meaningful, in virtue of having the status of a "grammatical rule," while (1) itself is treated as meaningless, in virtue of being the denial of a grammatical rule.[9]

I.2. Carnap. Carnap's neo-Humean empiricism provides a less radical response to Wittgenstein's problem. On its formal side, Carnap's philosophy is an explication of Gottlob Frege's semantics. The principal innovation in connection with analyticity is to replace Frege's vague remarks about definition with a formal means for specifying logical relations among extralogical terms. This enables Carnap to treat sentences like the denial of (1) and (5) as analytic within a calculus.

Carnap's[10] reasoning can be reconstructed as follows. Wittgenstein had supposed that the source of the contradiction in a sentence like (1) lies in the nature of the color words themselves, but that supposition is not forced on us. The contradiction does not have to arise from structure internal to the color words in order for it to be shown in a symbolism. After all, the logical property 'being a tautology' exhibited in instances of 'P v~P' is not shown in a representation of the internal structure of the logical words in those sentences. Rather, the property is shown in a symbolism of logical postulates that represents the external structure of logical words. So, if we take the contradiction in the case of (1)—and other semantic properties in other cases—to be genuinely logical properties like 'being a tautology', then we ought to treat so-called extralogical words in the same way we treat logical words. Therefore, the apparatus for logical postulates ought to be used to formulate a parallel set of "meaning postulates" that represents the external semantic structure of so-called extralogical words.

---

[9] At the outset of "the final stage," Wittgenstein diagnoses his earlier bewilderment as follows: "We make a picture like that of the two colors being in each other's way,...but on looking closer we find that we can't apply the picture we have made"—*The Blue and Brown Books* (Malden, MA: Blackwell, 1958), p. 56. Subsequently, the idea of "[a] picture [holding] us captive" becomes the general diagnosis of the metaphysical condition—*Philosophical Investigations* (Malden, MA: Blackwell, 1953), sections 115-32.

[10] "Meaning and Synonymy in Natural Language," *Philosophical Studies*, VII (1955): 33-47.

Since meaning postulates are constraints on the assignment of extensions to sentences, Carnap[11] can explicate analytic truth as truth in all possible worlds (*ibid.*, pp. 7-13). This explication has an ontological and an epistemological payoff. Since analyticity is necessary truth, the explication delivers logicism effortlessly and without Frege's commitment to mathematical realism, and since necessary truths have no factual content, it delivers empiricism without a commitment to John Stuart Mill's inductivist view of logical and mathematical knowledge. Reflecting on the significance of this explication, Carnap[12] wrote:

> What is important...is that it became possible for the first time to combine the basic tenets of empiricism with a satisfactory explanation of the nature of logic and mathematics. Previously, philosophers had only seen two alternative positions: either a non-empiricist conception, according to which knowledge in mathematics is based on pure intuition or pure reason, or the view held, e.g., by John Stuart Mill, that the theorems of logic and mathematics are just as much of an empirical nature as knowledge about observed events, a view which, although it preserved empiricism, was certainly unsatisfactory (*ibid.*, p. 47).

I.3. Quine. Quine's response to Wittgenstein's problem is part of his case for an empiricism in the Millian spirit. Although Quine[13] found Mill's conception of the certainty of logic and mathematics unacceptable, he found the Millian view of their empirical status congenial, and, accordingly, Carnap's empiricism seemed a compromise of the empiricist principle that all our knowledge rests on experience. Quine's[14] response to Carnap was clean and simple: extralogical words have no logical powers; hence there is nothing for

---

[11] *Meaning and Necessity* (Chicago: University Press, 1956, 2d ed.).

[12] "Intellectual Autobiography," in Paul A. Schilpp, ed., *The Philosophy of Rudolf Carnap* (La Salle, IL: Open Court, 1963).

[13] "Carnap and Logical Truth," in *The Ways of Paradox* (New York: Random House, 1965), pp. 100-01.

[14] *From a Logical Point of View* (Cambridge: Harvard, 1953).

Carnapian meaning postulates to explicate. Carnap's assumption, that extralogical words have logical powers, is "an unempirical dogma of empiricists" (*ibid.*, p. 37).

II. THE POSSIBILITY OF ANOTHER SOLUTION TO THE PROBLEM

The ways out of a problem divide into *solutions*, which typically reject an assumption which has been made about the problem but accept its presuppositions, and *dissolutions*, which reject one or another presupposition. Carnap's way out of Wittgenstein's problem is a solution. It rejects the assumption that logical properties and relations of sentences like (1)-(6) arise from the internal structure of their extralogical vocabulary. Wittgenstein's and Quine's ways out are dissolutions. They reject presuppositions. They both reject the idea that meanings are objects, that there is such a thing as necessity, that there can be a symbolism that shows that semantic properties are contained in the senses of sentences, and that our knowledge of the falsity of (1), the truth of (5), and the inconsistency of (2) and (6) is a priori. Here ends the similarity between their ways out. Quine's preserves the calculus approach of Frege/Russell style semantics and adopts a scientistic conception of philosophy; Wittgenstein rejects that approach and adopts a therapeutic conception of philosophy.

If these three ways out were exhaustive, the existence of a solution would depend on the adequacy of Carnap's way out. Hence the widespread belief that Carnap's way out is inadequate has led to the widespread belief that dissolution wins by default. But those three ways out are not exhaustive. The existence of another way out has been overlooked because philosophers equate intensionalism with Frege/Carnap intensionalism. I have argued elsewhere[15] that this equation is mistaken and that it limits our view of the options on a wide range of important philosophical issues. I shall argue here that the equation limits our view of the options on Wittgenstein's problem.

Quine was right about the inadequacy of meaning postulates: "We understand what expressions the rules attribute analyticity to, but we do not understand what the rules attribute to those expressions."[16] But the failure to provide such understanding is not, as Quine thought, due to there being *au fond* nothing for meaning postulates to explicate. It is due rather to the fact that meaning postulates are in their nature unsuited to explicating what is there. They derivationally enumerate sentences under the rubric *analytic*. The sentences are specified in the manner of theorems of logic, not as instances of the concept of analyticity. Thus, meaning postulates serve as constraints on the assignment of extensions to sentences, but they cannot explain the property common to the sentences they enumerate. Like Socrates's interlocutors, meaning postulates offer examples of the concept instead of the concept.

---

[15] *The Metaphysics of Meaning* (Cambridge: MIT, 1990); "The New Intensionalism," *Mind*, CI, 4 (1992): 689-720; "Names without Bearers," *Philosophical Review*, CIII, 1 (1994): 1-39; "Analyticity, Necessity, and the Epistemology of Semantics," *Philosophy and Phenomenological Research*, LVII, 1 (1997): 1-28; and *Sense, Reference, and Philosophy* (in preparation).

[16] *From a Logical Point of View*, p. 33.

This failure is evident in cases of co-extensive concepts. What is the meaning of the word 'two'? One answer is the postulate '$(x)(x$ is the number two $\leftrightarrow x$ is the successor of the number one$)$'. Another answer is the postulate '$(x)(x$ is the number two $\leftrightarrow x$ is the even prime$)$'. Still another answer is the postulate '$(x)(x$ is the number two $\leftrightarrow x$ is the square root of the number four$)$'. Indeed, any postulate in which the predicate in the consequent is necessarily co-extensive with 'two' will do, since Carnap's referential semantics provides us with no adequate basis for deciding among extensionally equivalent answers. Meaning postulates are incapable of distinguishing synonymous expressions from necessarily co-extensive expressions because they cannot provide the fine-grained notion of meaning required to explicate meaning in natural language. They cannot do this because, as representations of the external (that is, logical) structure of the extralogical words, they do not have access to the internal (that is, sense) structure of those words.[17]

Thus, a new way out must reject Carnap's assumption that the external, logical structure of extralogical words is the source of analyticity, contradiction, and analytic entailment in connection with sentences like (1)-(6). It must assume instead that such properties and relations derive from the internal, sense structure of extralogical words. The bottom line is this: if there is to be a solution, the notion of sense must be something other than Frege's and Carnap's referential notion of sense.

But does not the very definition of sense automatically make it a referential notion? Frege defined sense as the determiner of reference, and the subsequent Carnapian doctrine on which sense is a function from possible worlds to extensions is only a slight modification that brings Frege's definition in line with the modal expansion of the universe. But if there is a definition of sense that explains it in nonreferential terms, it might be possible to separate sense structure from logical structure, and this would open up the prospect of an alternative solution to Wittgenstein's problem.

---

[17] Carnap's notion of intensional isomorphism (*Meaning and Necessity*, pp. 56-64) and its progeny do not overcome the difficulty because they assume an isomorphism between sense structure and syntactic structure. I shall discuss this assumption below, see also my "Semantics in Linguistics and Philosophy: An Intensionalist Perspective," in Shalom Lappin, ed., *Handbook of Contemporary Semantic Theory* (Malden, MA: Blackwell, 1996), pp 612-13.

In the early seventies, I presented such a definition, namely, (D):

> (D) The sense of an expression is that aspect of its grammatical structure which is responsible for its sense properties and relations, that is, having a sense (meaningfulness), sameness of sense (synonymy), multiplicity of sense (ambiguity), repetition of sense (redundancy), opposition of sense (antonymy), and so on.[18]

(D) explicates sense not as the determiner of referential properties and relations like denotation and truth, but as the determiner of sense properties and relations like meaningfulness, synonymy, ambiguity, and redundancy. On (D), the theory of sense concerns the relations of senses with one another and sentences of a language, rather than, as in the Frege/Carnap tradition, with relations of the language to the world.

(D) enables us to make two important dichotomies. It enables us to separate sense structure sharply from both logical structure and syntactic structure. Sense is not only *not* something that determines reference, it is also *not* something that is determined by syntax. The former dichotomy frees us from interpreting sense (hence the notions of analyticity, contradiction, and analytic entailment) in logical terms. The latter frees us from the view of sense structure as isomorphic to syntactic structure. That view commits us to saying that syntactically simple words have simple senses, that their only semantic structure is external structure, and that meaning postulates are the best description of semantic structure. But if syntactically simple words can have complex senses, there may be a better description available to us.

I shall call the complex sense structure of syntactically simple words *decompositional sense structure.*[19] This level of grammatical structure introduces a different surface structure/deep structure distinction. Since compositional senses of expressions are a function of the senses of their constituents, the componential structure of such senses reflects the constituent structure of the expression. Since expressions with decompositional senses have no constituent structure, their componential structure is an intrinsic property of the expressions. Hence, as with cases where underlying syntactic structure cannot be read off surface structure (for example, 'Simpletons are easy to fool'), decompositional sense structure must be inferred from linguistic intuitions about underlying sense structure. We posit decompositional sense structure because only in this way can we explain sense properties and relations of expressions that cannot be explained otherwise. For example, only if we assume that the sense of 'bachelor' is complex can we explain the synonymy of 'bachelor' and 'single man'. The sense of 'single man' is complex, being a compositional function of the senses of 'single' and 'man'. Since 'single man' and 'bachelor' have the same sense, the sense of 'bachelor' is complex.

---

[18] *Semantic Theory* (New York: Harper and Row, 1972), pp. xvi-xxvii, 1-12.

[19] For a response to Wittgenstein's criticism of analysis, see my *The Metaphysics of Meaning*, pp. 52-66 and 101-07. For a response to the argument against decompositional semantics in Jerry A. Fodor, "Concepts: A Potboiler," *Cognition*, 1 (1994): 95-113, see my "Analyticity, Necessity, and the Epistemology of Semantics," pp. 13-16.

I shall argue shortly that the level of decompositional sense structure makes it possible to explain the semantic powers of logically elementary sentences on the basis of the nonlogical, internal structure in morphemes. Wittgenstein once took a step in this direction:

> ...from the given I can construct what is not given.... That makes it look as if a construction might be possible within the elementary proposition. That is to say, as if there were a construction in logic which didn't work by means of truth tables.
>
> What's more, it also seems that these constructions have an effect on one proposition's following logically from another.[30]

He went no further. The reason seems clear. A construction that explains the semantic powers of logically elementary sentences must be, as it were, both inside and outside of logic. The construction must be inside of logic to "have an effect on one proposition's following from another." It must be outside of logic to be "possible within the elementary proposition" and not to "work by means of truth tables." But with only Fregean definitions of sense and analyticity, the construction cannot be outside of logic.

We have just seen how, on (D), sense structure can be outside of logic: sense structure is a referentially independent linguistic structure that is internal to syntactically simple words in elementary sentences. In the next section, I shall spell this idea out more precisely. In section VII.2, I shall explain how it can also be inside logic in precisely Wittgenstein's sense.

---

[30] *Philosophical Remarks*, pp. 105-06.

III. TWO NOTIONS OF CONTAINMENT

Two conditions must be met for sense structure to be outside of logic. First, sense structure must be strictly mereological (part-whole) structure, and, second, the sense properties and relations must be defined in terms of sense structure (together with aspects of syntax). Logical operations can have no role in sense containment and hence in the definitions of analyticity, analytic entailment, and other sense properties and relations. The containment relation must be expressible solely in terms of the parts of senses and identity. For example, the redundancy of 'single bachelor' and 'male bachelor' must be definable as an identity of the sense of the modifier with a part of the sense of the head. Similarly, the analyticity of (10) and (11) and the analytic entailment of (8) and (9) by (7) must be definable in strict mereological terms:

(7) John is a bachelor.
(8) John is single.
(9) John is male.
(10) Bachelors are single.
(11) Bachelors are male.

The importance of this point is not the only reason for stressing it. Frege[11] muddied the waters when he introduced another sense of 'containment' in his definition of analyticity as provability from laws of logic and definitions (*ibid.*, p. 4). He was quite clear about the difference between the literal, mereological sense of 'containment' and his figurative sense. He likened the former, as it appears in Immanuel Kant's definition of analyticity, to the containment of beams in a house, and he likened the latter, logical notion, as it appears in his own definition, to the containment of a plant in the seed (*ibid.*, pp. 99-101). Just as a plant is not something there in the seed from which it grew, so, generally speaking, conclusions of logical inferences are not something literally there in the premises from which they are logically derived. This was a point on which Frege himself insisted (*ibid.*, p. 101).

The autonomy of the theory of sense depends not only on having a definition of sense like (D) but also, and as much, on having definitions of analyticity and analytic entailment based exclusively on literal containment. To be sure, Frege presented arguments against such definitions, but they are nowhere near as convincing as he took them to be. His principal argument was that they are not "fruitful" enough to do justice to definition in logic and mathematics (*ibid.*, pp.

---

[11] *The Foundations of Arithmetic* (Malden, MA: Blackwell, 1953).

99-101).[22] Kant's definition of analyticity does not, for example, even cover as simple a case as (12):

(12) If *A* is larger than *B* and *B* is larger than *C*, *A* is larger than *C*

The sense of the antecedent in (12) does not literally contain the sense of the consequent, since the terms '*A*' and '*C*' do not appear as arguments of the predicate 'larger than' in either conjunct of the antecedent. But so what if Frege is right that (12) and indefinitely many other more sophisticated sentences of logic and mathematics are not analytic in the Kantian sense? Since Kant did not think everything that is a priori and necessary is analytic, Frege begs the question.[23] Moreover, not everyone shares Frege's commitment to logicism or holds another doctrine that requires analyticity to cover mathematical and logical truth. And, of course, Frege does not address the question of whether the notions of definition and analyticity appropriate to natural language are fruitful in his sense.

Further, the fact that Frege was also right that Kant's subject-predicate definition of analyticity does not cover relational sentences is equally inconsequential because the failure is not due, as Frege would have it, to its failure to capture relational sentences like (12). The definition is incomplete because it fails to capture relational sentences like (13)-(16):

(13) Henry buys books from those who sell them to him.
(14) Jill walks with those with whom she strolls.
(15) Jack kills those he murders.
(16) Police follow those they chase.

The latter but not the former exhibit the same redundant predication found in sentences like (10) and (11). The only difference between analytic sentences like (13)-(16) and analytic sentences like

---

[22] Frege also says: "How can we [ask whether the predicate concept is contained in the subject concept] if the subject is an individual concept? Or if the judgment is an existential one" (*ibid.*, p. 100)? The first question is answered if Kant is allowed the same opportunity to invoke a descriptive view of names that Frege allows himself—see *Translations from the Philosophical Writings of Gottlob Frege*, Max Black and Peter Geach, eds. (Malden, MA. Blackwell, 1952), p 58 If such a view is false, as Millians think, Frege's question does not arise. The second question might be trouble for Kant, who argued against Descartes's ontological proof on the grounds that existential propositions are always synthetic. But this question is no trouble for a Kantian style account of analyticity per se, as argued in my account of Descartes's *Cogito*—see my *Cogitations* (New York Oxford, 1986), pp. 98-177.

[23] See Paul Benacerraf, "The Last Logicist," in P. A. French et alia, eds., *Midwest Studies in Philosophy, VI* (Minneapolis: Minnesota UP, 1981), p. 34, fn. 6.

(10) and (11) is that, in the relational sentences, some term other than the subject is the containing term. Thus nothing requires us to follow Frege in abandoning literal containment. All that is required is to abandon Kant's limitation of the containment condition to subject terms and to replace it with a generalization, something like (A):

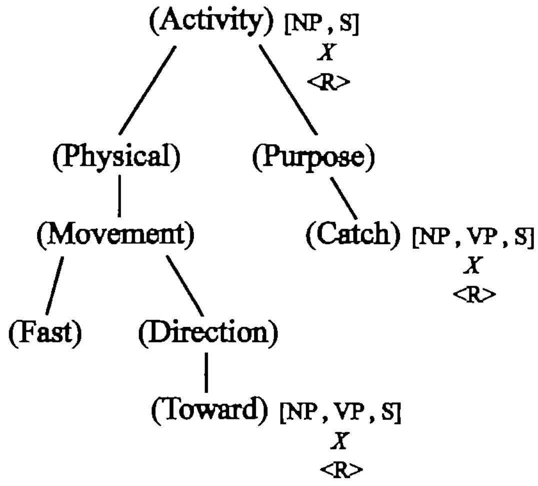
> (A) If the sense of a simple sentence consists of an *n*-place predicate with a term occupying each place, then the sense is analytic if one of the terms contains the full predicate with each of the terms.[24]

IV. REPRESENTING SENSE STRUCTURE

The principal task in this and the next two sections is to provide a symbolism that exhibits "the contradiction...contained in the *sense* of the two propositions." In earlier work, I developed a symbolism for sense structure (*ibid.*). The symbols representing sense structure, called *semantic markers*, can be written as tree structures. But unlike the domination relations in the familiar tree notation for phrase markers, which represent constituent/subconstituent relations in sentences, the domination relations in semantic markers represent sense/subsense relations in senses.

Let us suppose the tree structures (17) and (18) are semantic markers that represent, respectively, the senses of the verbs 'chase' and 'follow':

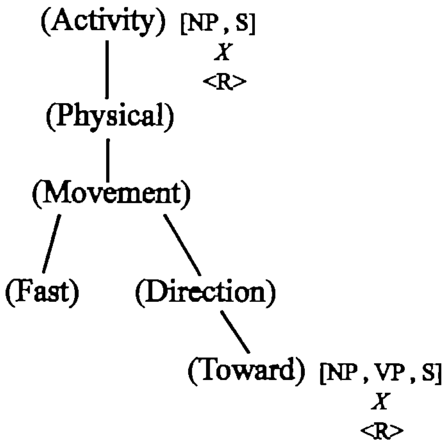
(17)



---

[24] *Semantic Theory*, pp. 174-77.

(18)



In terms of such a symbolism, we can define one sense to be the superordinate of another, its subordinate, just in case the latter is a same-rooted subtree of the former. Since (18) is a same-rooted subtree of (17), the semantic markers (17) and (18) represent the sense of 'follow' as a superordinate of the sense of 'chase'.

In explaining the compositional meaning of sentences, semantic markers combine in a way that represents the way in which the senses of the parts of syntactically complex constituents combine. One mode of marker combination is attachment of a semantic marker representing the sense of a modifier as a branch under the semantic marker representing the sense of its head. The compositional meaning of the expression 'follow with the purpose of catching the followee' could be obtained in this way, starting with (18) as the representation of the head and attaching a representation of the modifier to form (17).

Another way of obtaining semantic representations of compositional meanings is the embedding of one semantic marker within another. This form of marker combination takes place in the aspect of semantic markers that represents the argument structure of predicates. Occurrences of 'X' with syntactic information in brackets above the variable, such as in (17) and (18), are called *categorized variables*. They mark argument places. Note that the same categorized variable can occur more than once in a semantic marker, as in the case of (17). In this way, we can represent, inter alia, the fact that the sense of 'chase' identifies the object of the chaser's purpose with what directs the course of the chaser's activity. The placedness of a relation is, accordingly, specified by the number of distinct categorized variables that occur in its representation.

The syntactic information in the brackets specifies the semantic representations that can be values of the variable. It specifies the representations in terms of the syntactic constituents to which they are assigned.[25] For convenience, we may represent such information in the form of Noam Chomsky's[26] grammatical functions. They pick out a constituent on the basis of its syntactic role in the sentence: for example, the functions [NP, S] and [NP, VP, S] pick out, respectively, the subject of a sentence (the noun phrase that is a constituent of the whole sentence but of no other constituent in it) and the direct object of the sentence (the noun phrase that is a constituent of the verb phrase).

The angles under an occurrence of a variable state a condition that tells us what structure a semantic representation must have to be the value of the variable. These *selection restrictions* ensure that the results of attachment and embedding operations are bona fide semantic representations of senses. They are formal means to block what philosophers have called *category mistakes*. For example, the senses of the subject and direct object in (19) are of the appropriate category to combine with the sense of the verb 'chase', but the senses of the subject and direct object in (20) are not, as shown by the deviance of (20):

(19) The police chased the fugitive.
(20) The polygon chased the number seventeen.

Thus we want the selection restriction in (17) to require that the semantic marker '(Concrete)' appear as the label of the root in the representations for subjects and direct objects of 'chase'. Such a requirement blocks the semantic representations of the subject and direct object of (20) from being values of the variables in (17), and, consequently, (20) receives no semantic representation. In this way, selection restrictions enable us to mark the degree of ambiguity of expressions and sentences, that is, to assign *n* semantic representations to sentences with *n* senses.

A theory of decompositional and compositional meaning in the above sense consists of a dictionary in which each lexical item is associated with semantic markers representing its senses, a principle for assigning semantic markers from dictionary entries to occurrences of lexical items in sentences, rules for the compositional operations such as attachment and embedding, and a set of definitions of sense properties and relations stated in terms of formal features of semantic markers. With such a theory, the informal generalization (A) can be restated as (A*):

---

[25] The syntactic information in categorized variables thus ensures that the semantic representations of terms go in their proper argument place in a predicate (for example, so that 'John loves Mary' is not represented as synonymous with 'Mary loves John').

[26] *Aspects of the Theory of Syntax* (Cambridge: MIT, 1965).

> (A*) A semantic representation represents an analytic sense if it is of the form $P\{T_1,\ldots,T_n\}$, where $P\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ is the predicate and $T_1,\ldots,T_n$ are the terms occupying its argument places, and one of those terms, T_i, is of the form $P'\{T'_1,\ldots,T'_n\}$, and $P\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ is a same-rooted subtree of $P'\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$ and each term in $T_1,\ldots,T_{i-1},T_{i+1},\ldots,T_n$ is a same-rooted subtree of the corresponding term in $T'_1,\ldots,T'_n$n.

Given (A*) and the representations (17) and (18), we can say that (16) has an analytic sense.

Such a theory enables us to state Wittgenstein's methodological constraint formally, as (B):

> (B) A sentence's having a sense property or relation *K* is shown in a symbolism just in case a semantic representation of the sentence meets the definition of *K*.

Since a representation of (16) falls under (A*), (B) enables us to say that (16)'s analyticity is shown in our symbolism.

V. THE FULL DIMENSIONS OF THE PROBLEM

The good news is that a symbolism of the kind just sketched solves one aspect of the problem. The bad news is that it seems to put a complete solution beyond our reach.

First, the good news. W. E. Johnson[27] observed that "the understanding of the subject-term [in a "structural proposition" such as (5) expresses]...demands a reference of it to the general category there predicated of it," and that

> ...structural statements contain as their predicate some wide logical category, and their fundamental characteristic is that it is impossible to realize the meaning of the subject term without implicitly conceiving it under that category (*ibid.*, Volume II, p. 14).

Further, Johnson distinguished this relation from that of class membership, arguing that the relation of a determinate like 'red' to its determinable 'color' obtains in virtue of "a unique and peculiar kind of difference that subsists between the several determinates under the same determinable" (*ibid.*, Volume I, p. 176).[28] The special mode of difference that unites the determinates under a determinable is that no two of them can simultaneously characterize the same thing. Something can be simultaneously both red and square, but not simultaneously both red and blue.

---

[27] *Logic* (New York: Cambridge, 1921).

Further, a difference between the members of a pair of determinates under a determinable can be compared to a difference between the members of another pair under the same determinable, but such a comparison cannot be made across determinables. We can say that the difference between red and orange is less than the difference between red and blue, but we cannot sensibly say that the difference between red and orange is less than the difference between red and square. Finally, the determinates 'red', 'blue', and so on under the determinable 'color' are in themselves indefinable, since they have no internal structure in virtue of which they are color concepts.

Our formalism for semantic markers incorporates Johnson's insights. In particular, it incorporates his insight about the indefinability of color concepts.[29] We distinguish between primitive senses and simple senses. Primitive senses are not definable in the system of senses: the semantic markers representing primitives are not definable in terms of other markers. But primitive senses may be simple or complex: the semantic marker representing primitive senses may or may not have domination relations representing superordinate/subordinate structure.

This enables us to represent the senses of basic color terms as complex primitive senses.[30] Representing the senses of 'red' and

---

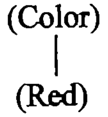
[28] See also Arthur N. Prior, "Determinables, Determinates, and Determinants," *Mind*, LVIII (1949): 1-20 and 178-94; and John R. Searle, "Determinables and Determinates," in Paul Edwards, ed., *The Encyclopedia of Philosophy*, Volumes 5 and 6 (New York: Macmillan, 1969), pp. 357-59.

[29] We may take G. E. Moore to be right about the indefinability of color terms—see *Principia Ethica* (New York: Cambridge, 1903), pp. 5-21. Hence the notion that for something to be yellow is for it to appear so to normal observers in normal circumstances is not to be construed as a definition of the term 'yellow', but as a condition for something to have the property. For an account of how a condition can be a condition for something to have a property without its being part of the definition of the term, see "Analyticity, Necessity, and the Epistemology of Semantics."

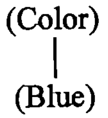
[30] We characterize basic color terms as those whose decompositional sense structure involves nothing more than the determinable 'color' with a single hue designation as its determinate. If the determinate is qualified in terms of degree of saturation or brilliance, or if more than one hue determinate occurs, the term is not a basic color term. Thus the lexical items 'red' and 'blue' (as represented in (27) and (28)) are basic color terms, while 'mauve', being the concept of a color of a bluish blue-red of high saturation and of medium to low brilliance, is not a basic color term.

'blue' with the markers (21) and (22) says that those senses are primitive but complex concepts:

(21)



(22)



Since the terminal components in these representations of the senses of 'red' and 'blue' have no status in the symbolism independent of their status in (21) and (22), (21) and (22) represent the intrinsic connection that the determinates 'red' and 'blue' have to their determinable 'color'.

We can now state some modest results. With the semantic marker '(Color)' as the semantic representation of the word 'color', (5) and similar sentences with other color words as their subject are analytic on the basis of (A*). The analytic entailment of (4) by (2) and similar analytic entailments can also be marked on the basis of a definition for analytic entailment parallel to (A*). In effect, the containment relation that (A*) requires $P\{T_1,...,T_n\}$ to bear to a term T_1 is the relation that the representation of an entailed sense must bear to the representation of the entailing sense.[31] This solves one aspect of the problem. Given (B), the analyticity of the sentence (5) and the analytic entailment of the sentence (4) by the sentence (2) are shown as "contained in the *sense* of the...propositions."

Now for the bad news. The first sign of trouble comes when we try to use the symbolism thus far set out to account for the analyticity of a sentence like (23):

(23) Red is not blue.

Since the function of the English words appearing in semantic markers are purely mnemonic, semantic markers like (21) and (22) are simply distinct symbols for distinct senses. Thus they do not distinguish antonymous terms like 'red' and 'blue' from merely nonsynonymous terms like 'spouse' and 'parent'. (21) and (22) must be appropriately modified not only to make this distinction, but also to

[31] *Semantic Theory*, pp. 171-97.

enable us to represent (23) as analytic in the sense of (A*). But this means that the sense of the subject 'red' in (23) must literally contain the sense of the predicate 'not blue'. This is hard to imagine, but, to make matters worse, the sentences in (24) have to be represented as analytic as well, and this means that the senses of 'not green', 'not yellow', 'not brown', and so on must likewise be packed into the sense of the subject 'red':

(24) Red is not green, Red is not yellow, Red is not brown,....

Hence it appears as if the sense of every basic color term must somehow have the senses of all the other basic color terms as parts. Here it looks very much like our system, too, "begins to crack."

Once, not so long ago, only a tiny minority of philosophers, notably Arthur Pap and William and Martha Kneale,[illegible] held the view that sentences like (23) are synthetic. Now it is the majority view. Not surprisingly, the arguments for this view are variants on the theme that the sense of a basic color word cannot contain the sense of another basic color word. For example, C. L. Hardin[illegible] says:

> There need be no occasion for a reflective green-seer to have had contact with red,...[hence] there is no reason to suppose that the concept of not being red is thereby part of her concept of being green; after all, not being a mastodon isn't part of the concept of being a lampshade, even though reflective people would be readily able to distinguish mastodons from lampshades (*ibid.*, p. 122).

At this point, the full dimensions of the problem are visible. Since literal containment provides the only alternative to Carnap's unacceptable way out of Wittgenstein's problem, it is essential for a solution. But literal containment appears to rule out a solution in the test case.

VI. HOW TO DO THE SEEMINGLY IMPOSSIBLE

I shall now show that, appearances to the contrary, (23) is analytic. To do this, we have to provide a compositional derivation of (23) which begins with representations of the senses of the lexical items in the sentence and ends with a representation of its sense on which the sense of its subject 'red' contains the sense of its predicate 'not blue'. The account of the derivation must indicate how, *mutatis mutandis*, we can give similar derivations for the other color sentences in question.

[illegible] Pap, *Elements of Analytic Philosophy* (New York: Macmillan, 1949), p 422; and Kneale and Kneale, *The Development of Logic* (New York: Oxford, 1962), p. 637.

[illegible] *Color for Philosophers: Unweaving the Rainbow* (Indianapolis: Hackett, 1986).

VI.1. Taking stock. Before turning to this derivation, we need to take stock of what we already know about the senses of the items 'red', 'be', 'not', and 'blue'. In an earlier work,[34] I developed an account of the sense concept of negation. This account takes off from the distinction between external and internal negation drawn by Arthur Prior[35] and others. Their distinction is based on scope. The scope of external negation is the whole sentence, while the scope of internal negation is a phrase or smaller constituent. External negation is, of course, defined extensionally, in terms of truth and falsehood. But internal negation is defined extensionally, too, in terms of set-theoretic relations among extensions of terms. Looking at internal negation through the lens of Fregean sense, intensionalists see another form of logical negation.

(D) allows us to carry the sharp sense/reference distinction we have drawn for the rest of the vocabulary through to the negative forms of the language. We are thus committed to defining internal negation *intensionally*. We can define it as an operator that converts senses in its scope into other senses. We may illustrate this with the antonyms in (25):

(25) perfect, flawed

'Not perfect' or 'imperfect' means 'flawed', and 'not flawed' or 'unflawed' means 'perfect'. In this case, internal negation functions as an operation taking us from one sense to another, from one member of an antonymous *n*-tuple to the other. Here, the difference between the external negation of standard logic and the internal negation of language stands out clearly. The former is a toggle, a toggle between truth values. The latter is a toggle, too, but a toggle between senses.

Given that internal negation is a sense toggle, the internal negation of a sense is some other sense which—as internal negation in connection with (25) illustrates—belongs to the same antonymous *n*-tuple. Since each of the members of an antonymous *n*-tuple have a common superordinate but stand in an exclusion relation to each other member, the internal negation of a sense is some one of its co-members. The sense of an internal negation is thus an operator, which we symbolize as "A/...," whose application to a sense yields an exclusive disjunction of the other senses in its antonymous *n*-tuple.

We already know some things about the senses of 'red' and 'blue'. One is that the hue concepts in those senses are determinates of the determinable 'color'. This means that (21) is part of the semantic representation of 'red' and (22) part of the semantic representation of 'blue'. We also know that those semantic markers are not full representations of 'red' and 'blue', since we know that (21) and (22) do not represent the antonymy of 'red' and 'blue'. The markers are all right as far as they go, but they do not go far enough.

---

[34] *Semantic Theory*, pp. 157-71.

[35] "Negation," in Edwards, Volumes 5 and 6, pp. 458-63

Interestingly, we know something about the further parts of the markers that represent the antonymy of 'red' and 'blue'. Since (26) raises the same question for us as (23), the parts of the semantic markers for the senses of 'red' and 'blue' that represent their antonymy relations must be the same modulo the difference between (21) and (22):

(26) Blue is not red.

Moreover, since there is a set of sentences for (26) corresponding to (24), that is, 'Blue is not green', and so on, the semantic representation of a basic color term must somehow show its incompatibility with all other basic color terms. Hence the decompositional sense of a basic color term has to embody the antonymy structure of the entire system of basic color terms.

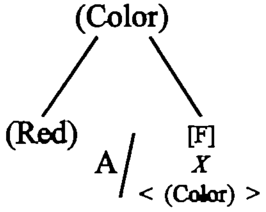
Here we come to the key point. A basic color term must express its incompatibility with all the other basic color terms without the senses of the latter themselves occurring in the sense of the former. The element of truth in the claim of Hardin and others is that one color concept is not part of another. Their mistake was to conclude from this that sentences like (23) are not analytic on the basis of the same containment relation found in (10) and (11). They overlook the possibility that the antonymy structure of the color words is embodied in their senses in a sufficiently abstract form for both this element of truth to be preserved and such sentences to be analytic.

VI.2. The derivation. To represent the antonymy structure of color words as abstractly embodied in their senses, the semantic markers (21) and (22) have to contain a further branch, Br, that expresses the antonymy of the color concept they represent with the other color concepts, but does so without semantic markers for them appearing in Br. Since Br is the same for the representation of every color word, we need only figure out what it is in the case of (21). To do this, let us think ahead to the critical step in the derivation at which Br is transformed into a branch Br* that represents the antonymy of the sense 'red' with the particular color concept 'blue'.[36] The semantic marker for the sense of 'blue' that enters Br at this step can only come from the representation of the sense of the word 'blue' in the verb phrase of (23). Since, as (24) shows, any color word could appear in place of 'blue' in this verb phrase, Br must contain a categorized variable at which to embed a representation of the sense of whatever color term we find in the verb phrase. This means that Br is a branch whose topmost marker '(Color)' dominates a complex symbol consisting of an occurrence of 'A/...' associated with a categorized variable whose value is a semantic representation of the constituent of the verb phrase containing the color term. Hence (27) replaces (21) as the semantic representation of the lexical item 'red':

---

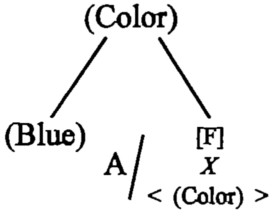
[36] There must be such a step if the derivation is to provide a representation of (23) on which it is analytic.

(27)



The syntactic information in [F] and the selection restriction together tell us that values of that variable are semantic representations of the predicate in the verb phrase just in case the superordinate marker in the semantic representation of the adjective is '(Color)'. Adding the Br to (22) provides (28) as the lexical representation of 'blue':

(28)



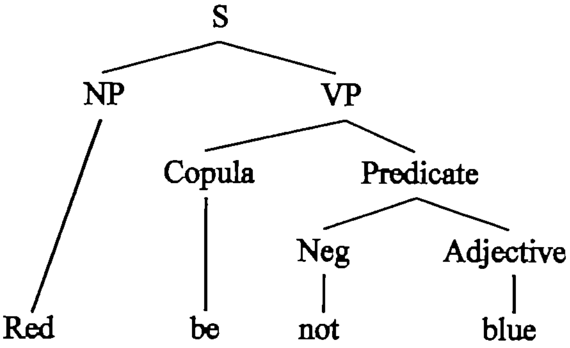
The copula 'be' is the one lexical item in (23) still requiring a representation. We introduce the schematic marker (29):

(29) [Predicate, S] [NP, S]
X X
< > < >

It is modeled on the familiar notion of the application of the predicate concept in a subject-predicate sentence to the subject concept. The grammatical function '[Predicate, S]', which is read as "the constituent dominated by 'Predicate', dominated by 'S'," picks out the predicate 'not blue' in (23). The function '[NP, S]', as indicated earlier, picks out the subject of the sentence. (29) becomes the representation of (23) when the variable categorized '[Predicate, S]' and the one categorized '[NP, S]', respectively, are replaced with the representations of the predicate and the subject of (23).

Since the derivation of a representation of (23) describes its compositional meaning, the representation is a function of the aforementioned representations of 'red', 'be', 'not', and 'blue', and the syntactic relations among the constituents of (23). Those relations can be depicted in the phrase marker (30) for (23):[27]

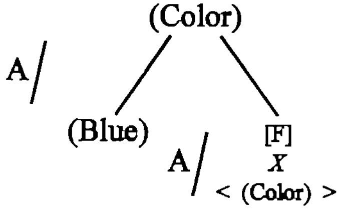
(30)



Since compositional principles work up from the most deeply embedded constituents—the lexical items of the sentence—to the least deeply embedded constituent—the entire sentence—the first step in our derivation assigns semantic markers from the dictionary to the terminal symbols of (30). At this step, (27) is assigned to 'red', (28) to 'blue', 'A/...' to 'not', and (29) to the copula 'be'.

The second step combines the representation of 'not' with the representation of the adjective 'blue' to get (31) and assigns (31) as the representation of the predicate 'not blue' in (30):

(31)

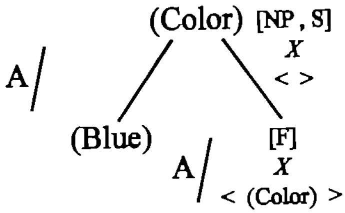


---

[27] The syntactic formalism in the text, though dated, is preferable for present purposes to the formalism of current versions of syntactic theory. There are too many of the latter, they change often, philosophers are less familiar with them, and they are less transparent on first encounter.

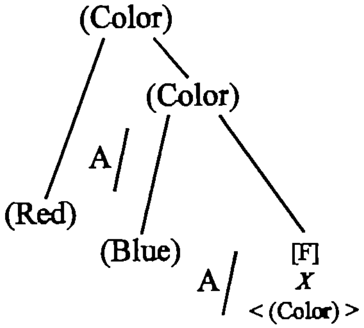
The third step makes (31), just assigned as the semantic representation of the predicate, the value of the variable in the semantic marker (29) that is categorized for representations of predicates. The result is (32), which is assigned as the semantic representation of the VP constituent in (30):

(32)



The fourth step provides a value for the categorized variable in the representation of the subject in (30). Since the semantic representation of the predicate, (31), has the marker '(Color)' as its topmost marker, it satisfies the selection restriction of the variable in (27), and becomes its value. The result, (33), is assigned to the subject in (30):

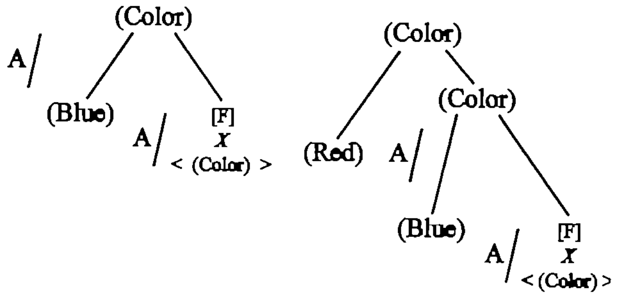
(33)



In the final step of the derivation, (33), which was just assigned as the representation of the subject in (30), is made the value of the variable categorized for the subject in (32), and the result, (34), is assigned to the sentence constituent in (30), and hence is the semantic representation of (23):[illegible]

[illegible] The occurrences of the categorized variables do not remain in (34). When there is no semantic representation to provide the value of a variable, its value is the semantic marker in its own selection restriction. This principle is independently required to handle analytic entailments, such as that from 'The dog drank' to 'The dog drank a liquid'—see *Semantic Theory*, p. 107.

(34)



VI.3. Analyticity and contradictoriness. (34) is of the form $P\{T_1,...,T_n\}$, $n \geq 1$. The semantic marker on the left of (34) which represents the predicate $P\{x_1,...,x_n\}$ is a same-rooted subtree of the semantic marker on the right which represents the term T_i, and hence (A*) marks the sense represented by (34) as analytic.

Similarly, sentences like (1) and (35) can be marked as contradictory on the basis of representations like (27) and (28):

(35) Red is not red.

On the definition of contradictoriness, a sense of a sentence is contradictory if either there is an antonymy among its predicates or there is an antonymy between its predicates and its terms (*ibid.*, pp. 180-81). (1) is an example of the former type of contradictoriness, and (35) is an example of the latter. In the case of (1), (27) and (28) are each values of the occurrence of the variable in the other. When these saturated semantic markers become the value of the variable categorized '[Predicate, S]' in the occurrence of (29) assigned to the copula of (1), the sentence will be represented as having an antonymy relation between its component predicates, and so be marked contradictory. In the case of (35), each occurrence of (27) will be the value of the variable in the other occurrence. When these saturated semantic markers replace the variables categorized '[Predicate, S]' and '[NP, S]' in the occurrence of (29) assigned to the copula of (35), the result is a semantic marker that represents (35) as having an antonymy between its predicate and term. Thus (35) will be marked contradictory.

VII THE REMAINDER OF THE SOLUTION

Since, for us, analyticity and contradictoriness are relations internal to the senses of sentences of a language, not relations between sentences and the world, section VI is not yet a complete solution to Wittgenstein's problem. We have still to explain why and in what sense analytic sentences like (23) are necessary truths and contradictory sentences like (1) are necessary falsehoods.

VII.1. The explanation. I shall not argue that analytic sentences are straightforward necessary truths. The topic of whether they are is better left for another time.[39] In any case, such a construal of them is not open to us, since it would treat the senses of the sentences in question as logical forms and analyticity as a species of logical truth. Fortunately, Saul Kripke's[40] notion of weakly necessary truth does just as well for present purposes as necessary truth: "We can count statements as necessary if whenever the objects mentioned therein exist, the statement would be true" (*ibid.*, p. 137). In this spirit, let us adopt the definitions (WNT) and (WNF):

> (WNT) A proposition *p* is a *weakly necessary truth* if and only if *p* contains no inconsistency among its terms and *p* is true in all possible worlds where each of the terms in *p* occurring in a referential position has a nonempty extension.[41]
>
> (WNF) A proposition *p* is a *weakly necessary falsehood* if and only if *p* contains no inconsistency among its terms and *p* is false in all possible worlds where each of the terms in *p* occurring in a referential position has a nonempty extension.

If a sense of a sentence is analytic, it is a weakly necessary truth. The reasoning behind this is as follows.[42] Given *p* is analytic in the sense of (A*), *p* has a term *t* that contains the entire sense *p*. If there is no inconsistency among the terms in *p*, there are worlds in which there are things of which *p* is true or false. Since the truth conditions for *p* are included in the conditions for the extension of *t*, *p* is true in all those worlds. Hence *p* is a weakly necessary truth. Since (23) is analytic in the sense of (A*) and has no inconsistency among its terms, it is a weakly necessary truth. Thus, being red is necessarily connected with not being blue.

---

[39] To show this requires construing simple sentences like (10) and (11) as material conditionals to deal with worlds in which there are no bachelors or colors for such sentences to be true of. Construing simple subject-predicate sentences as compound sentences was Frege's device for, inter alia, bringing subject-predicate sentences under his definition of analyticity. Such construals, as I argue in *Sense, Reference, and Philosophy*, are dubious linguistics, because they distort the grammar of simple sentences. As David Wiggins and others have pointed out, universal sentences like (10) are not complex sentences formed with the sentential connective 'if, then', and particular sentences like 'Some bachelors are single' are not complex sentences formed with the sentential connective 'and'—"'Most' and 'All': Some Comments on a Familiar Programme, and on the Logical Form of Quantified Sentences," in Mark Platts, ed., *Reference, Truth, and Reality* (New York: Routledge, 1980), pp. 318-46.

[40] "Identity and Necessity," in Milton Munitz, ed., *Identity and Individuation*, (New York: University Press, 1971).

[41] For an explanation of the notion of a referential position that appears in the definitions (WNT) and (WNF), see my "A Solution to the Projection Problem for Presupposition," in Choon-Kyu Oh and D. A. Dinneen, eds., *Syntax and Semantics, Volume 11: Presupposition* (New York: Academic, 1979), pp. 91-126.

[42] "Analyticity, Necessity, and the Epistemology of Semantics," pp. 24-25.

If the sense of a sentence is contradictory, then it is a weakly necessary falsehood. Since (1) and (35) are contradictory and evaluated only on worlds where there is something for them to be true or false of, they are weakly necessary falsehoods. Thus, 'red' and 'blue' are necessarily exclusive and 'red' and 'not red' are necessarily exclusive.[illegible]

VII.2. The notion 'inside logic'. As we saw at the end of section II, a solution to Wittgenstein's problem requires that senses be both outside and inside logic. I explained above how senses can be outside logic. I shall now explain how they can also be inside logic. Sense is inside of logic in that senses play an essential role in the connection between sentences and logical properties like weakly necessary truth and weakly necessary falsehood. As we just saw, the sense structures underlying analyticity, contradictoriness, and analytic entailment are, inter alia, part of the explanation of why (5) and (23) are weakly necessary truths and why (1) and (35) are weakly necessary falsehoods. It is easy to see that sense structure plays the same role in explaining why inferences like those from (2) to (3) and (2) to (4) are valid. Since senses provide the fine-grained linguistic structure necessary for a model-theoretic explanation of why such sentences have such logical properties and relations, senses are inside logic in precisely Wittgenstein's sense of "hav[ing] an effect on one proposition's following from another."

VIII. WHY THE SEEMINGLY IMPOSSIBLE SEEMED IMPOSSIBLE

Why did it seem impossible for the sense of 'red' in (23) to contain the sense of 'not blue'? The answer that emerges from our discussion is that no distinction was made between the meaning of a word *in isolation* and the meaning of a word *in a sentence*. To be sure, it is impossible for the sense of the lexical item 'red' to contain the sense 'not blue', but the case of a word in a sentence is different. It is different because the sense of a word in a sentence can have sense components which are not part of its sense as a lexical item but come

---

[illegible] Weakly necessary truth is not only the referential correlate for analytic sentences, it is also the correlate for denials of contradictory sentences. Hence the sentence 'It is not the case that the spot is red and blue' is a weakly necessary truth. Correspondingly, weakly necessary falsehood is not only the referential correlate of contradictory sentences, it is also the referential correlate for sentences that are denials of analytic sentences. Hence the sentence 'It is not the case that red is not blue' is a weakly necessary falsehood.

into the latter from the senses of other constituents in the sentence. Compositionality makes the difference. Failing to note this, philosophers assimilate the case of words in a sentence to the case of words in isolation, thereby taking a genuine impossibility intuition about the latter to apply to the former, where it does not apply at all.

Restricted to lexical concepts, Hardin's claim that "the concept of not being red is...[not] part of [the] concept of being green" expresses an uncontroversial truth. But such claims are intended to apply generally, and they must do so in order for them to support the view that Wittgenstein's problem has no solution. But if the claims do apply generally, they apply to senses of words in sentences, and then, as we have seen, they are false.[44]

IX WHY THERE IS NOT A CONFLICT BETWEEN INTUITION AND LOGIC

Philosophers saw Wittgenstein's problem as posing a conflict between intuition and logic. It is true that sentences like (1)-(6) have the semantic properties and relations that intuition tells us they have. It is also true that the components of (1) are elementary propositions, that (2), (4), (5), and (6) are themselves elementary propositions, and that (3) is the negation of an elementary proposition, and hence it is true that those sentences cannot have their semantic properties and relations on the basis of their logical structure. But there is no inconsistency between intuition and logic because the semantic structure on the basis of which the sentences have those properties and relations is not logical structure. The alleged conflict between intuition and logic disappears once we stop thinking that semantics, in both senses of this ambiguous term, is logic.[45]

---

[44] Hardin is correct to claim "not being a mastodon isn't part of the concept of being a lampshade." But Hardin's example 'A lampshade is not a mastodon' is hardly a parallel case to (23). Hardin's example is not analytic, while, as we have seen, (23) is. The difference lies in the fact that the sense of 'red' and the sense of 'blue' are members of the same antonymous *n*-tuple, whereas the sense of 'lampshade' and the sense of 'mastodon' are not members of the same antonymous *n*-tuple, but only contain senses (that is, the senses 'artifactual' and 'natural') that are. This difference allows the sense of 'mastodon', the predicate in Hardin's example, to contain senses that do not appear in the sense of 'lampshade', the subject in that example. Thus Hardin's example is not analytic. But, although it is not, it is a weakly necessary truth because it is the denial of the contradictory sentence 'A lampshade is a mastodon', and being a weakly necessary truth is also the referential correlate of being the denial of a contradictory sentence.

[45] This is not an isolated case The same thinking underlies a number of other alleged conflicts: for example, the conflict between Millian and descriptivist views of names and the conflict between internalist and externalist views of meaning. See my "Names without Bearers," *Philosophical Review*, CIII, 1 (1994): 1-39; and B Armour-Garb and my "What's in a Name" (to appear) regarding the former, and "Analyticity, Necessity, and the Epistemology of Semantics" regarding the latter.

X. A METAPHYSICAL CODA

The failure of Frege/Carnap's semantics was presumed to show that dissolution is the only way out of the problem. Since dissolution, of both the Wittgensteinian and Quinean type, would leave the rationalist with no basis for claiming that knowledge of color incompatibility is a priori, it is seen as removing a major obstacle to naturalized epistemology and hence as a success for naturalism and empiricism.[46] This perception has led to a shift in the focus of contemporary philosophy from efforts to explain modal facts about color incompatibility to efforts to explain why intuition misled us and why it is so difficult to find a plausible example of something simultaneously red and blue.[47] Naturalists and empiricists are now pursuing approaches to such explanations, which, at one extreme, represent a return to something like Husserlian phenomenology, and, at the other, represent a return to something like Spencerian biologism.[48] The present paper opposes this claim of a success for naturalism and empiricism.

Initially, the paper might seem to support empiricism. That is because a central theme of logical empiricism was that the statements about color that we have been considering are analytic a priori truths. By exhibiting "the contradiction [as] contained in the *sense* of the two propositions," our symbolism provides a solution on which knowledge of color incompatibility is indeed analytic a priori knowledge. But the notion of analyticity that our argument justifies is not the Frege/Carnap notion that expands the class of analytic truths to encompass the necessary truths. Instead, it is the mereological notion which contracts that class so that virtually all logical and mathematical truths are synthetic. Thus, acceptance of our solution to the color incompatibility problem entails, for the empiricist, the undesirable consequence that the problem of synthetic a priori knowledge reemerges for logical and mathematical truths generally.

---

[46] See, for example, Arthur C. Danto's remarks in Hardin's *Color for Philosophers: Unweaving the Rainbow*, pp. x-xii, and Hardin's remarks, pp. 121-27.

[47] Danto (*ibid.*, p. xii) engages in a bit of rationalist bashing on the basis of Hewitt Crane and Thomas P. Piantanida's interpretation—"On Seeing Reddish Green and Yellowish Blue," *Science*, ccxxi (1983): 1078-80—of the reports of some of their subjects as to what they see when looking at the point where a field of red and a field of green merge. The reports are supposed to demonstrate that sentences like (2) and (6) can both be true in application to the same spot. Quite apart from the conflation of the cases Hardin calls "conjoining" and "compounding," this jumps the gun. As Hardin himself observes, there was no replication of the experiment as of the time of his writing and, as he also observes, the experiment has been "greeted with widespread skepticism in the visual-scientific community" (*ibid.*, p. 125). Moreover, why take the interpretation of the reports of the subjects in question at face value? It is not as if the responses of the subjects in the experiment were uniform, or as if one could not come up with other explanations of the responses of the subjects in question. Taking their responses, or similar responses of people in various homespun experiments, as veridical reports of a homogeneously red and green surface is as misguided as taking comparable reports about figures in the area where a series of polygons merge into a circle as establishing that there are figures which are simultaneously a polygon and a circle.

[48] On the phenomenological side, there is Colin McGinn's claim that sentences like the negation of (1) express "necessities of how the world can seem in perceptual experience"—*The Subjective View* (New York: Oxford, 1983), pp. 21-28. McGinn's phenomenology does not solve Wittgenstein's problem because a solution to that problem is under the modal constraint of showing that the negation of (1) is necessarily false in the Leibnizian sense of holding in no possible world. On the Spencerian side, there is Quine's approval of evolutionary explanations of certain "structural traits of color perception"—Quine, *Ontological Relativity and Other Essays* (New York: Columbia, 1969), p. 90.

JERROLD J. KATZ

Graduate Center/City University of New York

COMMENTS AND CRITICISM

PRESTON ON EXAPTATION: HERONS, APPLES, AND EGGS

Beth Preston's[1] attempt to exploit the concept of exaptation in her proposed theory of function puts a salutary strain on it, thereby exposing its weaknesses better than its critics have done. Stephen Jay Gould and Elizabeth S. Vrba[2] once proposed a contrast between adaptations and what they called *exaptations*: "characters evolved for other usages (or for no function at all) and later 'coopted' for their current role" (*ibid.*, p. 6). The term never caught on among evolutionary biologists (I can find only one use of the term by Gould himself outside the original article[3]), and Ruth Millikan[4] and I have independently argued that exaptation is nothing but the early stage of any adaptation, rather than a distinguished phenomenon in its own right. As I put it:

> every adaptation is one sort of exaptation or the other—this is trivial, since no function is eternal; if you go back far enough, you will find that every adaptation has developed out of predecessor structures each of which either had some other use or no use at all.[5]

Preston claims that Millikan and I both make the same mistake: we "neglect to ask whether there are any ongoing exaptations that do not get transformed into adaptations [in the way just described]...and so enjoy an independent status" (242 fn. 31). Let us see, then, if there are exaptations that are not just the juvenile stages—as one might say—of adaptations.

Preston: "The case of the mantling heron shows that there are indeed ongoing exaptations" (241). African black herons use their out-

---

[1] "Why Is a Wing Like a Spoon? A Pluralist Theory of Function," this JOURNAL, XCV, 5 (May 1998): 215-54.

[2] "Exaptation: A Missing Term in the Science of Form," *Paleobiology*, VIII (1982): 4-15. The paper is often cited by philosophers, and is reprinted in David L. Hull and Michael Ruse, eds., *The Philosophy of Biology* (New York: Oxford, 1998).

[3] *Bully for Brontosaurus. Reflections on Natural History* (New York: Norton, 1991), p. 144n "Gould and Vrba offer the term 'exaptation' for such traits, but even though their paper was published 16 years ago the term has found little use in biology"—R. McN. Alexander, "Finding Purpose in Life" (review of Colin Allen, Marc Bekoff, and George V. Lauder, eds., *Nature's Purposes: Analyses of Function and Design in Biology*), *Science*, CCLXXXI, 14 (August 14, 1998): 927

[4] *Language, Thought, and Other Biological Categories* (Cambridge. MIT, 1984); *White Queen Psychology and Other Essays for Alice* (Cambridge. MIT, 1993).

[5] *Darwin's Dangerous Idea: Evolution and the Meanings of Life* (New York: Simon and Schuster, 1995), p. 281.

0022-362X/98/9511/576–80

spread wings to shade the shallow water in which they wade, cutting the reflective glare and permitting them to see their prey. The example is drawn from Gould and Vrba's paper, but instead of providing the support she supposes, it actually undermines her reading of their concept. As Gould and Vrba point out, there is a genetic basis for the mantling behavior pattern. So, although the shape of the wings may not (yet) have been significantly adjusted to this new use, there is no doubt that the use has been shaped by selection pressure, and hence that mantling is an adaptation after all. Preston surmises that "the fact that only one species of heron does do it indicates that the selection pressures favoring it are not very strong. Since wings of birds that do not fly rapidly become vestigial, the pressure would have to be strong to maintain wings suitable for mantling in a flightless heron" (240-41). This hypothetical speculation is both dubious and beside the point. Compare the prospects of herons that can mantle but not fly with herons that cannot do either; the selection pressure for mantling would be very strong indeed in that imagined scenario, and the fact that the behavior is found in only one species of heron is no indication at all of weakness of selection pressure. But in any case, since multiple-selection pressures act simultaneously, and since the mantling behavior is clearly under selection, we need not attempt the thankless task of apportioning the relative contribution of flight and mantling (and other yet unnoticed functions) to the maintainance of the shape (and opacity, and weight, and so on) of the wings.

All adaptations start with some fortuitous aptness, a lot or a little, which then gets further refined (a little or a lot) to serve the new function better. If we want, we can honor the cases at one extreme with a term of their own—we can call *exaptations* those cases in which there is lots of initial aptness (by somebody's lights) and relatively little if any later adjustment (so far). Gould and Vrba observe that the *objets trouvés* that form the basis of adaptations are very often exaptations in this sense, and this is important, since it alerts us to the omnipresent possibility that features of adaptations may owe more to selection pressure under an earlier selectional regime than to any selectional shaping for current use. This salutary reminder may well serve to correct a variety of temporal myopia that sometimes afflicts adaptationists, but does not mark a major challenge or alternative to adaptationism, as Gould and Vrba suggest.

Preston tries to make good on their suggestion. She sees that, if exaptation is to be something that contrasts with adaptation, rather than just an early phase of adaptation, she needs to find cases which are not evanescent and in which the aptness of a feature for a function is *clearly* independent of any "ongoing" selection pressure shaping or maintaining it. Gould and Vrba do not offer any such examples, nor do they harbor any such ambitions for their exaptations. One can see why. So long as the genetic variation for an exapted feature shares a genome with the genetic variation for whatever feature underlies the "exapting," the selection pressures determining their respective fates are inevitably intertwined, dissolving the imagined barrier between exaptation and adaptation. This becomes clear if we imagine varying the case of the mantling heron so that the chances for the sought-for independence are maximized. Suppose herons mantled by holding a large, broad leaf in their bills, flipping it aside at the last instant as they plunged their bills into the water. The shape of the leaf so used would be under selection pressure in the heron genome (shaping the leaf-choosing machinery, or even the leaf-shape-altering machinery, in the heron's brain), while the original shape of the leaf would be presumably beyond the reach of the heron's genome, under "independent" selection pressure to serve other ends in the plant's genome. I flag the presumption of independence, however, since even here, with two entirely distinct genomes, there is the omnipresent possibility of interaction effects: such herons would presumably favor fishing in locales with good shade leaves, and might even enter a symbiotic partnership with the favored plants, providing some benefit in exchange for leaves, thereby getting in position to have a selectional effect on the original leaf shape *for mantling*![6]

Exaptation-without-adaptation can thus be unstable even when the phenomenon is split between two different species. But other interspecific cases might provide the sort of independence Preston seeks. Consider a difference between apples and eggs. Both are mighty good eating. Apples are adaptations, naturally selected *for* being good to eat, in payment to frugivores for spreading seeds. (Cultivated apples enhance this adaptation, of course.) Eggs are not adaptations for being good to eat; there is no selection pressure for deliciousness in eggs. (Domesticated hen eggs are an interesting exception, of course, since they *have* been under artificial selection pressure for excellence as food.) Even wild eggs are delicious and nutritious; they are apt for eating, as many ovivores attest. Haunch of antelope is also delicious and nutritious, I gather, and certainly lions often select it as food, but it was not designed to be food even though lions and other carnivores have been designed to use it as food. What if we were to consider eggs and antelope haunches as exaptations? This is clearly not what Gould and Vrba had in mind, however ubiquitous such aptnesses are in nature. They are better for Preston's purposes, though, since they are parallels, in biology, for the "standardized exaptations of artifacts" Preston proposes in her pluralistic theory. All living things are food for some other living things, and *food* is certainly a functional category (like *medicine*, *lubricant*, and the like). Yet living things and their parts are not in general designed by natural selection to be food.

---

[6] This sort of interaction is often overlooked, but its significance for evolutionary theory is the main message of Richard Dawkins's most important book, *The Extended Phenotype* (San Francisco: Freeman, 1982), which proposes and defends a new "central theorem": "*An animal's behaviour tends to maximize the survival of the genes 'for' that behaviour, whether or not those genes happen to be in the body of the particular animal performing it*" (p. 233).

Nature is full of items which are not artifacts at all but which nevertheless are prized by one organism or another for performing some function. The gravel in a hen's gizzard is not designed for its grinding role, or for any other role; hens may be quite picky, though, about what bits of gravel end up in their gizzards. Or consider salt, which plays such a life-enhancing role in the lives of us all, to say nothing of the air we breathe (try living without it) and the very ground we stand on, for that matter. If hemoglobin molecules have the function of transporting oxygen molecules around in the blood stream, what is the function of the oxygen molecules? These items maintain their aptness for their functional roles independently of any shaping by natural selection of *their* properties; natural selection designs the surrounding systems to take advantage of them. How shall we apply the concept of exaptation to this ubiquitous feature of evolution? Shall we call all these cases exaptations? If it were not for Preston's attempt to keep Gould and Vrba's concept of exaptation from evaporating under scrutiny, these problems with it might have gone unnoticed.[7]

[7] In a long footnote, Preston claims that I make "a rather elementary mistake in moving from 'every adaptation developed out of an exaptation' to 'every adaptation is an exaptation'" As she points out: "We do not say that because every bird developed out of an egg, every bird is an egg" (243 fn. 31). True indeed. I make a similar point in *Darwin's Dangerous Idea* (p. 206) But in the context in which I make the claim (quoted above), my meaning is clear: every adaptation is an exaptation at some point in its career, just as every bird is an egg at some point in its career Contrary to what Preston says, I do not "repeat the slogan that all adaptations are exaptations at intervals," though it is true that the many adaptations I discuss all start out as exaptations, as I sometimes note. I do say (once, p. 390) that all exaptations are adaptations, and Preston attempts to rebut this claim as well, by finding me (and Millikan) guilty of a fatal equivocation: "in the case of exaptation, 'selec-

Preston is right to deplore the paucity of work on artifacts by philosophers, but the lacuna is not quite so gaping as she supposes. Some, but not all, of the good points in Preston's essay are made by me in my various discussions of "artifact hermeneutics."[8]

DANIEL C. DENNETT

Tufts University

---

tion' means adoption for a particular use on a particular occasion regardless of whether that use is successful or is repeated" (243 fn. 31). I doubt if this is what Preston meant to say, since it licenses a profligacy of function beyond all imaginable theoretical interest. On this definition of exaptation, "selection" is so leniently defined that every stone a creature stepped on would become a stepping stone, even if the creature stumbled on it. Since this was probably not her intention, her objection to Millikan and me remains unclear.

[8] "Beyond Belief," in A. Woodfield, ed., *Thought and Object: Essays on Intentionality* (New York: Oxford, 1982); "Evolution, Error, and Intentionality," in *The Intentional Stance* (Cambridge: MIT, 1987); and "The Interpretation of Texts, People, and Other Artifacts," *Philosophy and Phenomenological Research*, L, Supplement (1990). 177-94.

BOOK REVIEWS

The Significance of Free Will. ROBERT KANE. New York: Oxford University Press, 1996. 268 p. Cloth $48.00.

This is an admirably ambitious and thoroughly engaging book. Robert Kane's thesis is that free will, "the power to be an ultimate creator and sustainer of one's own ends or purposes," is "a significant kind of freedom worth wanting" (15) which is incompatible with determinism but consistent with a naturalistic view of human beings and their world. Kane defends the claims about incompatibility and significance in Part I of the book, where he also offers a naturalistic account of the will (chapter 2). His aim in Part II is to advance an account of free will that can hold its own against the familiar charge that libertarian freedom is "essentially mysterious" or "terminally obscure" (13). On Kane's naturalistic view, the will operates causally, but indeterministically, in a way that has a "place in the modern picture of human beings that is emerging in the natural, social, and cognitive sciences" (115). In the context of this emerging picture, he also speculates about how free will may be realized in the brain.

As I say on the book's dust jacket, this is, "quite simply, the most thoughtful and detailed defense of libertarianism currently available." Kane sheds light on why "alternative possibilities" are important to libertarians, productively shifting the focus of the debate to the deeper issue of "ultimate responsibility." Far from being content to attack compatibilism, he offers a detailed defense of the intelligibility and possibility of libertarian free will that features probabilistic causation and shuns mystery. The book's first chapter masterfully motivates the free-will issue, and, throughout, Kane offers a superb critical review both of traditional work on free will and of the recent literature.

Kane argues that "[a] willed action is 'up to the agent' in the sense required by free will only if the agent is ultimately responsible for it" (35). In basic instances of ultimate responsibility, there is an internal conflict—for example, "between what an agent believes ought to be done and what the agent wants or desires to do" (126). If agents' choices "are not determined in such cases," Kane writes, they "might choose either way" (127). "The choice in moral and prudential conflict situations terminates an effort (to resist temptation) in one way or another" (127), and since the effort is "indeterminate," "the

0022-362X/98/9511/581–84

choice that terminates it is *undetermined*" (128). A failed effort terminates in a choice of the tempting course of action, whereas a successful one more happily terminates in a choice of the subjectively right course. "[T]he agent's past character and motives...influence the choice without determining it" (127). "Some of the agent's reasons or motives (the moral or prudential ones) explain why the agent makes an effort to resist temptation, while others (self-interested motives or desires for present satisfactions) explain why it is such an *effort*" (128).

Libertarians rightly see determinism as a bar to agents' having causally open alternatives (that is, alternatives consistent with their past and the laws of nature), and they traditionally maintain, as does Kane, that such alternatives are required for free will and moral responsibility. Indeterminism removes this bar, but it also raises a worry about *luck*. If there are causally undetermined or indeterminate aspects of a process that terminates in, for example, a choice and these aspects are present at the very time the choice is made and directly relevant to the process's outcome, then, to the extent that the agent is not in control of these aspects, luck enters the picture in a way that threatens both moral responsibility and a desirable freedom. Luck is problematic when it seems significantly to impede agents' control over themselves. Given the indeterminacy of an agent's effort to resist temptation on Kane's view, it may seem that luck plays too great a role in the success or failure of the effort for the agent to have freedom-level control over the outcome and to be morally responsible for it. An agent who luckily succeeds might have unluckily failed; in some nearby possible worlds, that is precisely what happens.

Although at one point Kane claims that in conflict scenarios of the sort I have described, agents who make causally undetermined choices "might choose either way, *all past circumstances remaining the same* up to the moment of choice" (127; emphasis added), he takes this back when responding to a particular worry about luck. Kane argues that, given the indeterminacy of efforts to resist temptation, we cannot "imagine the same agent in two possible worlds with exactly the same pasts making exactly the same effort and getting lucky in one world and not the other. Exact sameness or difference of possible worlds is not defined if the possible world contains indeterminate efforts or indeterminate events of any kinds" (172).

Consider the following story. John believes that he ought to arrive on time for a meeting that begins at noon in his building, but he is tempted to arrive late, as a modest protest. Although he tries very hard and very intelligently to resist his temptation, and although he has a reasonable chance of succeeding, his indeterminate effort fails: a minute before noon (at *t*) he decides to go to the meeting late. At a nearby world with the same laws and a very similar past relative to *t*—as similar as can be regarding John (or his counterpart), given the relevant indeterminacy—John* (that is, John or his counterpart) believes that he ought to arrive on time for the meeting, is tempted to arrive late (for the same reason), and tries very hard and very intelligently to resist his temptation. John*'s indeterminate effort succeeds. He masters his temptation, decides a minute before noon to show up on time for the meeting, and does so.

John's effort to resist temptation fails whereas John*'s succeeds. If there is nothing about the agents' powers, capacities, states of mind, moral character, and the like that explains this difference in outcome, then the difference is just a matter of luck. If John and John* do not differ in the control they exert over their respective efforts, or differ in control only in ways that are themselves a matter of luck, John's succumbing to temptation whereas John* successfully resists is a matter of luck. That their efforts are indeterminate explains why the efforts might have different outcomes, but this obviously does not explain (even in an indeterministic way) why John failed whereas John* succeeded. If it were not for John's having worse luck than John*, John would have been in John*'s shoes: he would successfully have resisted temptation. Given that John tried very hard and intelligently to resist temptation and that his failing where John* succeeded is just a matter of John's being unluckier than John*, it is difficult to see why John should be thought to have exerted freedom-level control in making the decision to go to the meeting late or to be morally responsible for making that decision.[1]

I turn to the significance of free will. In a chapter on that topic, Kane writes: "when one traces the desires we have for incompatibilist free will to their roots, by way of [ultimate responsibility], one eventually arrives at two elemental...desires—(i) the desire to be independent sources of activity in the world...and (ii) the desire that some of our deeds and accomplishments...have objective worth" (98). He invites us to "understand what objective worth is" by considering a pair of stories about an artist named Alan (97). In the first, a benevolent friend deceives Alan into believing that "he [Alan] is a great artist, and...is being duly recognized as such." In the second, "Alan has many of the same experiences, including the belief that he is a great artist," and what he believes is true. In both stories, "Alan dies happily, believing he is a great artist." Kane remarks that "to say that there is an important difference in value in the two [cases] for Alan, even though he would not know it and would *feel* equally happy in both, is to say he endorses a notion of *objective* worth" (97), and he suggests that "the notion of ultimate responsibility is of a piece with this notion of objective worth" (97-98).

---

[1] I develop this worry in "Ultimate Responsibility," *Social Philosophy and Policy* (forthcoming, 1999).

This is unpersuasive. Even if Alan is a compatibilist and has no desire for ultimate responsibility in Kane's indeterministic sense, he can reasonably prefer the second scenario to the first. A compatibilist can reasonably prefer actually being a great artist and having the associated feelings and experiences of success to having the same feelings and experiences while falsely believing himself to be a great artist. He can prefer that these qualitative states are grounded in his excellent work and that he not be radically deceived about himself.

This leaves it open, of course, that the significance of libertarian free will rests on the other desire Kane emphasizes, "the desire to be independent sources of activity in the world." Even so, one would like to be confident that this desire can be satisfied in an indeterministic agent without its satisfaction's hinging on the agent's being subject to a kind of luck that precludes moral responsibility and any kind of free will worth wanting. I am doubtful that the primary location Kane has selected for freedom-promoting indeterminism (and indeterminacy) in choice-producing processes will prove to be adequate for the libertarian's purposes. But even if this location is problematic, Kane has made remarkable progress in developing libertarianism, and perhaps some alternative location will fare better. This book is certainly on the cutting edge, and it is already shaping the current debate on free will. Moreover, given Kane's lively, accessible prose, it is ideal for use in courses on the topic, including undergraduate courses.

ALFRED R. MELE

Davidson College

Achieving Our Country: Leftist Thought in Twentieth-century America. RICHARD RORTY. Cambridge: Harvard University Press, 1998. 159 p. Cloth $18.95.

This is the first of Richard Rorty's books to take up a challenge he has put to himself, ever since *Philosophy and the Mirror of Nature*,[1] about the imperative need in America for a public philosophy in the manner of John Dewey. An educated public can be trusted to recognize the broad character of social problems. But people will be edified, Rorty has been saying, only if the choices in view are well analyzed—by philosophers among others; above all, by thinkers who are exposed to the crosscurrents of a society, conversant with its history, and reasonably sympathetic to its ideals. Exposure, in the sense of vulnerability and inquisitiveness, may be the hardest condition for an academic to satisfy today. University life in itself makes a limited milieu: we are as perfectly informed and perfectly insulated as the White House press corps. Rorty anyway sees the depth of the problem. He knows what they are saying outside the universities, and his title strikes a patriotic note that has long been detectable in his political comments—especially in the 1980s, when he risked his cachet with the academic left by varying his characteristic tone (good natured and quotable, often at the brink of dandyism) to deal out occasional thunderbolts against the Soviet empire. Anticommunism, and a distrust of the post-Sixties style of jeering at America and all its works, may still be views that separate him from a fair portion of the academic left. Nevertheless, *Achieving Our Country* is a lay sermon chiefly addressed to them. It comes out of three Massey Lectures, delivered at Harvard University in 1997, where Rorty began by gloomily reflecting that "the only version of national pride encouraged by American popular culture is a simpleminded militaristic chauvinism" (4). In a way, it would be heartening if this were true; it would offer both an inviting and a familiar target for reformers; but the description has surely become less true in the years between the Reagan and Clinton presidencies. The current clichés of the media picture America as a center of multicultural blending, the peaceful conqueror of the global market. How do you fight against that?

A good deal hangs on what we make of pride. Friedrich Nietzsche remarked in *Daybreak* that the moral world after God would require

---

[1] Princeton: University Press, 1979.

0022-362X/98/9511/585–90

"more pride" or "a different pride"—he was not sure which. A similar perplexity seems to have troubled Rorty when he wondered about the fate of equality and decency after the decline of national sentiment on the American left. Should one appeal to more of the old pride, or a different kind of pride? Rhetorical as well as pragmatic considerations come into play here. If you can get people to like the good things better with American names, then why not call the good things American? All the more when there is a historical warrant for it, and when, in keeping with the broad-church communitarianism of our time, you feel compelled to assign all values a local meaning in a particular language game. And yet, Dewey and Walt Whitman—the other hero of *Achieving Our Country*—are dubious allies on this point; when Rorty says "they wanted Americans to take pride in what America might, all by itself and by its own lights, make of itself" (16), his supporting quotation from Whitman actually says something different and appeals to a different entity: "And I call to mankind...." Can we split the difference then by agreeing that "Both Dewey and Whitman viewed the United States as an opportunity to see ultimate significance in a finite, human, historical project, rather than in something eternal and nonhuman" (17)? But this series of contrasts (finite-historical-human versus eternal-nonhuman; local source versus "ultimate significance") reflects Rorty's preoccupations more than Whitman's or even Dewey's. For moral and political reasons, he does not want to praise America too locally, or uniquely; while, for philosophical reasons, he has to be sure not to align himself with the eternal or the more-than-historical. This brings up a curious point about his stance. Rorty would by no means wish to be called a philosopher's philosopher; yet he has never been content to work things out like a nonphilosopher's nonphilosopher. His arguments naturally cast themselves in the shapes of the discipline, and he always borders them with position-taking markers, even in work that speaks explicitly against the necessity of doing so. The dualisms of philosophical discussion are immensely real to him, far more than they are to the ordinary person, and when he puts his hand on his heart and says about the good of free speech or equal rights, "This I believe, and can in consistency believe, without believing in the correspondence theory of truth," even the lay reader who can see what he is saying will wonder why he wanted to say it like that. The allusions to a philosophical story, which take up much of the first chapter and a bit of the third, are a superfluous trimming unless one supposes that many literate Americans suffer from an excess of faith in objective truth.

Like all of Rorty's books, this one has a recountable plot. Until the coming of Stalinism, he thinks, the American left had an impressive unity based on distributive justice, the eradication of prejudice, and the reduction of the privilege of wealth and status. But in the 1930s, the Communist Party commandeered the native democratic project, and thereafter, so long as the cold war was on, liberal anticommunism was the true home of the reformist left. Liberalism suffered a breakdown, however, under the pressure of the war in Vietnam. Meanwhile the New Left followed its own course of self-creation and self-destruction. The Students for a Democratic Society went from the "participatory democracy" of the Port Huron statement to the Maoism of the early 1970s, and the student left generally came to despise what remained of the old left. There were political causes of this rejection which Rorty does not mention: the adherence, for example, of right-wing social democrats to the most violent abuses of cold-war policy (the AFL-CIO under George Meany never allowed itself to criticize Lyndon Johnson's bombing of Vietnam), and the moral weakness of the arguments for "guns and butter" advanced by the circle around the League for Industrial Democracy, a group largely populated by ex-Trotskyists. But the student left acted from a moralized politics in which it lacked the self-irony to perceive the selfish element, and the mass movement it forged was undermined by an extraordinary lack of historical knowledge and curiosity, and by a double standard which favored the victim, however murderous, over the member of the oppressor class, however innocent. In the face of the lasting divisions that followed, Rorty concludes with an ecumenical appeal:

> As I see it, the honors should be evenly divided between the older, reformist Left and the New Left of the Sixties. The heirs of that older Left should stop reminding themselves of the stupid and self-destructive things the New Left did and said toward the end of that decade. Those who are nostalgic for the Sixties should stop reminding themselves that Schlesinger lied about the Bay of Pigs and that Hook voted for Nixon (71).

As the last two references make plain, *left* is a term that widens here to cover liberal and centrist reformers; and, if anything, Rorty errs on the side of tenderness in his treatment of the New Left. He thinks the war in Vietnam could not have been stopped by the tactics of older reformers alone, and quotes with approval a cheerful euphemism of Paul Berman's about "the slightly crazy attempt to raise insubordination into a culture" (70). The final handshake makes for a warming picture, but it leaves out one detail: from Tom Hayden's speeches to the contents of the Black Panther newspaper, the New Left after 1968 was not just hostile to the anticommunist left, but trusting of the totalitarian left. *Revolution* in those years was a word of strict and unambiguous praise. The word and the idea, vaguely entertained, became a fetish whose only possible rivals were "the people," "the streets," and "the children." Some of the worst of the Stalinist left survived in the New Left, and some of the worst of the New Left has survived in the academic left. Rorty is aware of these affiliations, but is not obliged to say all that he sees in a book written with the aim that this book has, and for the audience it hopes to create. He has never been a good hater.

Essentially, the argument is that the academic left, which he calls the "cultural left," should pick itself up, look abroad into the world, and stop condescending to ordinary persons whose chances in life are sinking, persons not sanctified by the glamor of race or gender deprivation. It is a sound recommendation, with a precedent in writings by Russell Jacoby, John Patrick Diggins, and others, but the advice will not be followed and the reasons are worth counting. With the displacement of the New Left by its successor, left-yuppie culture, the disenfranchised victim has had to take on more rarefied forms, just as the privilege of the sympathetic spectator has become more firmly entrenched. We have seen in our time the development of a culture in which *moral perfection is prized as one more commodity*, possession of which infallibly adds to the lustre of personal success. Nor need the attainment be vouched for by personal practice. The passive white middle-class supporter of affirmative action may nurse a scorn for the conservative black educator or businessman for which it would be hard to find any precedent in history or morality. Meanwhile, and it is the great achievement of the relativism of the educated middle classes, the one precept a child is sure to learn thoroughly in the first six years of school is that "everybody has their own opinion." What is good for multicultural education is good for the big corporations, too: advertising is one more way of making and holding opinions; and short of a war nobody wants, it can only hope to conquer the world by putting up a multiracial front. The versatility of this tissue of justifications—which keeps the poor of all races poor indeed—is wholly compatible with political quietism.

The entertainment-information complex is proving much harder to fight than the military-industrial complex. It does not rely on chauvinism, but a hybrid of personal and commercial pride; and nobody, including Rorty in this book, has yet analyzed how the democratic surface of the phenomenon may have changed the logic of social reform. Before being exposed to any political or philosophical culture, young people now are taught to construct their self-images by personalized consumption, and to regard the world as an extension of desired or undesired self-imagery. In this way of thinking, Dewey's "public" easily becomes an excluded term, and perhaps an unnecessary one. "Our country" may be the next to go: the global market is sacred, the nation and the people who constitute it secondary and expendable. And here one comes up against the stubbornness in Rorty's choice of terms—a stubbornness that is part of his integrity; and yet it stops him from thinking through the consequences of certain perceptions. A steady but not a predictable feature of his social criticism has been his disdain for Theodor W. Adorno (whom he seems to judge mainly by the *Dialectic of Enlightenment*) and for American writers like Christopher Lasch with roots in the Frankfurt School. Rorty's pragmatism is to some extent antipsychological, and he is reluctant to think that capitalism, or any other system of life, can deeply change people's habits of mind. If one supposed it could, then the change itself would be a subject for analysis, because it presents an obstacle to reform. As a writer on politics, Rorty draws the line at terms of argument that William James and Dewey would have understood. And yet "liberal consumerism," as William Finnegan has called it, may be a new thing in the world; if so, its ideology will be a problem facing the next generation of reformers—because it makes consumption so cool and liberated an option, and because it offers a simulacrum of liberalism. In the habitual pairing of "freedom" and "free market," it is the latter that, now comes first in almost every public setting. The priority no longer emanates from Chicago economists and libertarian fanatics. One can see it daily on the *New York Times* editorial page; and it was conspicuous, and scarcely remarked by the press, in President Clinton's speeches in China (Spring 1998), with their careful language promoting markets first and hoping that political liberty might follow as a supplementary blessing. This is an aspect of the public and its problems that may have changed qualitatively since Dewey's time.

Rorty reprints at the end two separate short essays: one on the virtues of "campaigns" as against "movements," the other on "the inspirational value of great works of literature." Maybe only a member of a literature department can fully appreciate the élan of the second piece, as argument and as gesture, so overgrown that it seems almost wild. It took courage and wit to sing out this reminder against the social construction of literature on a front-line panel at the Modern Language Association. Regarding the essay in praise of campaigns one may feel more reserved. It is certainly a piece of consistent pragmatism—in line with James's preference for "ambulatory" over "saltatory" arguments—and Rorty has an optimistic motive. Great modern projects like socialism were movements, he says, but they have ended now and we must get over our nostalgia for them. He comforts us by recalling how ambiguous the legacy was: the cityscapes of Le Corbusier and Mies van der Rohe were not quite meant for human use; and for the common purposes of social life, the same may be true of the brilliant prospectuses of Karl Marx and Friedrich Engels. We are so sure of how little we know about large-scale reform; campaigns flatly seem to have the better claim. But again, the matter looks different from the point of view of psychology and human action. People get warm enough to stick with a cause—marching or lobbying or speaking on its behalf—only when they feel that something big is at stake: a long-term good that connects with some deep passion, and not a middle-term benefit that connects with personal prudence. That is why the last movement of the left was the anti-war movement of the Sixties. Feminism and all the culturalisms, under the separate banners of whose glory it is impossible to launch a unified protest on a single issue, with the result that only hundreds rather than tens of thousands marched against the repeal of welfare: these are the fruits of the culture of campaigns, which has already triumphed beyond any need of a philosophical defense.

DAVID BROMWICH

Yale University

Nietzsche's Philosophy of the Eternal Recurrence of the Same. KARL LÖWITH. Translated from the German by J. Harvey Lomax; foreword by Bernd Magnus. Berkeley: University of California Press, 1997. xxviii + 276 p. Cloth $40.00.

J. Harvey Lomax offers us the first ever English translation of what both he and Bernd Magnus, to whom we owe the foreword, describe as an enduring "classic" of Nietzschean studies. First published in 1935 in Berlin, the book earns this distinction, apparently, on two grounds: it articulates a comprehensive original interpretation of Friedrich Nietzsche's philosophy, centered around the concept of nihilism and the doctrine of the eternal recurrence, and it exposes a fundamental tension in the latter doctrine.

Although the book does have these merits, it is also, by the current standards of philosophical writing, written obscurely: it satisfies itself too frequently with mere repetitions and paraphrases of Nietzsche's cryptic metaphors, and it falls short on detailed examination of the philosophical substance of his texts. As a result, even the basic task of sorting out the broad lines of the book's argument proves unnecessarily difficult. Nevertheless, the overall shape of Karl Löwith's interpretation is not uninspired and deserves our attention.

The interpretation developed in the book differs significantly from the prevalent trends of current Nietzschean scholarship: it makes no mention at all of views and concepts that are central to recent prominent interpretations, such as "perspectivism"[1] and the "will to power."[2] It plays down the critique of truth which animates the former and ignores the metaphysics associated with the latter. Instead, it emphasizes what might be called Nietzsche's "existentialism," namely, his concern with the value or meaning of our existence. According to Löwith, the foremost problem of Nietzsche's philosophy is nihilism, and the doctrine of the eternal recurrence is his proposed solution to it.

In the most general terms, nihilism is the view that our life, particularly the suffering in it, is meaningless. The modern European nihilism Nietzsche confronts results from the demise of a world view whose paradigm is Christianity. This world view expresses a complex

---

[1] For example, Alexander Nehamas, *Nietzsche: Life as Literature* (Cambridge: Harvard, 1985); Maudemarie Clark, *Nietzsche on Truth and Philosophy* (New York: Cambridge, 1990).

[2] For example, John Richardson, *Nietzsche's System* (New York: Oxford, 1996).

0022-362X/98/9511/591–97

fundamental attitude Nietzsche calls the "negation of life," apparently for two reasons. First, it rests on the judgment that our life in this world is worthless: it is a "wrong road" or a "mistake."[3] According to the Christian version of this view, our very existence is essentially sinful, and the suffering which afflicts us must be understood as punishment. Second, this world view also assigns a specific purpose to our life in the here-and-now: to prepare, through systematic ascetic self-denial, for another, better life in the "divine world."

This world view unravels with the "death of God," the realization that our life in this world is *all* there is, that there is no eternal "divine" life outside of it to redeem it. Löwith points out that the import of this realization is deeply ambiguous. On the one hand, God represents "the biggest objection against existence" (47); the death of God, accordingly, restores "innocence" to our existence in this world. On the other hand, the belief in God also gives our life in this world a meaning: it is a "bridge" to another existence. As a consequence, the death of God deprives our existence of its meaning (49).

Which of these possible consequences of the death of God one emphasizes, and so how one understands the significance of the nihilism it precipitates, depends on one's point of view (50ff.). Until now, the point of view of "weakness" has prevailed, a Nietzschean concept which Löwith explains inadequately but which points to another crucial dimension of the idea of the "negation of life": nothing in our life in this world can redeem the sufferings it causes; its justification must therefore be found outside of it. From this point of view, the death of God is a disaster which leads to "suicidal nihilism," or a "will to the nothing" (53). But there might be another point of view, from which the death of God opens up radically new possibilities and so constitutes a call for a new practical stance—a new "*responsibility* for the future" (85). If our life in this world is indeed all there is, we should no longer look at it as a "bridge," a transitory state which will find its redemption in an illusory "beyond"—its justification must be found in the here-and-now. This is the point of view of the "affirmation of life."

How can our life in this world be justified without appeal to an eternal life outside of it? Nietzsche's answer is the doctrine of the eternal recurrence: our life is fully justified if it is such that we can will its eternal recurrence. The doctrine is intended to fill the void left by the demise of the Christian myth of the eternal life. Löwith does remark on the analogies Nietzsche himself draws between his own views and Christian doctrine. For example, he notes Nietzsche's claim that *Thus Spoke Zarathustra*, where he introduces the idea of eternal recurrence, should be read as the "fifth Gospel" (61-62); and he gestures toward, but fails to examine closely, the relation between Nietzsche's idea and the Christian notion of the eternal life. Such an examination is critical, as it should explain why Nietzsche's doctrine takes on its peculiar form. One comparison worth sketching out here concerns the relation of the eternity of a life to its worth.

---

[3] "The idea at issue here is the *valuation* the ascetic priest places on our life: he juxtaposes it (along with what pertains to it: 'nature', 'world', the whole sphere of becoming and transitoriness) with a quite different mode of existence which it opposes and excludes, *unless* it turn against itself, *deny itself*: in that case, the case of the ascetic life, life counts as a bridge to that other mode of existence. The ascetic treats life as a wrong road...or a mistake that is put right by deeds..."—*On the Genealogy of Morals*, Walter Kaufman, trans. (New York: Vintage, 1989), III, 11.

In Christian doctrine, the relation between the eternity of the eternal life and its worth is complex. On the one hand, the eternal life cannot be worth living simply because it is eternal—because, for example, it is a life in which death has been vanquished: there is, after all, such a thing as eternal damnation. The value of the eternity of the eternal life depends directly on the value of this life itself: it must possess a kind of value that would make it worth living forever. On the other hand, even though the eternal life is not good simply because it is eternal, it is not fully good *unless* it is eternal. The finitude of a life that is otherwise good still counts as an objection against it.

Nietzsche rejects the second aspect of the Christian doctrine, presumably because it is an expression of the negation of life. If life is not fully good unless it is eternal, then our life in this world, which is essentially finite, cannot find its justification within itself. It is, in this respect at least, radically unredeemable. But he remains inspired by the first aspect of the Christian doctrine: it is only because a life in the presence of God possesses a certain kind of value that we are prepared to will the eternity or, one may presumably say, the eternal repetition of this life. In other words, the desirability of its eternity is a criterion of the value of a life. Nietzsche's doctrine of the eternal recurrence is nothing other than a proposal to apply this criterion to the justification of our life in this world: it will be justified if it is such that we can will its eternal recurrence. So understood, the doctrine of the eternal recurrence places a new *"ethical imperative"* upon us: "live in every moment so that you *could* will that moment back over and over~~" (85).~~

This first interpretation of the doctrine of the eternal recurrence brings us to the critical core of Löwith's book. He points out that Nietzsche presents his doctrine not just as an "ethical imperative" but also as a *"scientific theory,"* according to which our life will actually recur for all eternity and, indeed, has already occurred: "human life is turned upside down again and again like an hourglass" (88). The two versions of the doctrine appear to be in direct contradiction: as an ethical imperative, it presupposes that our future is open and to be determined by an act of free will (86), whereas as a scientific theory, it claims that our future is fated, and implies that such willing is pointless: "how can one...will something that excludes and renders superfluous, through the irrevocability of its fatality, every willing..." (87)?

The distinctive feature of Lōwith's interpretation is to suggest that Nietzsche was aware of this difficulty and to maintain that we should therefore understand the doctrine of the eternal recurrence not as the willing of an open future in the here-and-now, but as a reconciliation with fate, or "*amor fati*" (74-75). Although Nietzsche characterizes *amor fati* in his early writings as "freely willed necessity" (126), his considered view is, according to Lōwith, at once more subtle and less paradoxical:

> 'Loving' the absolute or fatal necessity is no longer a willing but...a condition in which the will no longer wills anything.... The accident of being-there—the accident that is deprived of its innocence by the belief in willed, purposeful creation of Being out of the nothing—is 'redeemed' in *amor fati*, because Zarathustra-Dionysus grasps precisely in accident what in the whole is as it must be (79-80).

Unfortunately, Lōwith's proposal is poorly laid out and what can be made out is fraught with difficulties. The idea seems to be the following: so long as we regard our existence as something that could have been different or not at all ("the accident of being-there" which results from "the willed, purposeful creation of being out of the nothing"), then there is a need for a justification of this fact. If there was nothing first, and it took an act of will for something to come out of it, then we want to know for what "purpose" something was willed into existence. With the death of God, the customary justification is no longer available. Recognizing the implications of the eternal recurrence as a scientific theory—that the world, and our existence in it, is "as it must be"—appears to solve the problem of justification by eliminating the "absolute contingency" of our existence (188): we become able to "grasp precisely in accident what in the whole is as it must be."

If this is Lōwith's proposal, it is fatally flawed, for natural necessity does not make for ethical justification: the fact that our life could not have been different does not make it worth living. It makes it impossible for us not to live out our fate, but it gives us no reason to do so. Löwith invites us to distinguish between two ways of overcoming nihilism, however: either we might create new "goals and purposes" for our life to replace the old Christian ones and justify our will to live; or we somehow jettison the very need for goals, and thus cease willing altogether. According to Löwith, the latter is how we must understand the notion of *amor fati*.

The natural necessity of our existence is not meant to justify our willing it as it is, it is meant to undermine willing. Our existence runs its fated course inexorably, utterly indifferent to the "goals and purposes" we might posit to justify it. A full appreciation of this fact is bound to discourage the positing of such goals and leave us in a state in which "the will no longer wills anything"—"a freedom from all goals and purposes, from every for-the-sake-of" (57).

If it is understood as the *imperative* to free ourselves from all goals and purposes, this proposal might still seem susceptible to the objection it was meant to disarm. For as a kind of imperative, the call for *amor fati* presupposes that our future is open and to be determined by an act of free will which the doctrine of the eternal recurrence understood as a "scientific theory" simply excludes. But Löwith subtly remarks that *amor fati* cannot be, by its very nature, the object of an imperative: it "can no longer be the goal of a willing but can only be an event and a metamorphosis that happens of itself" (80). The loving acceptance of one's fate cannot be a matter of compliance with an imperative.

It is unclear, however, how this observation is supposed to circumvent the objection. To say that the "freedom from all goals and purposes" cannot be "the goal of a willing" is not in the present case to concede that it is a fatality beyond our deliberate control. Rather, it points to a peculiar feature of the aspiration to this sort of freedom: we cannot free ourselves from the willing of goals by willing that as a goal, and so it would be incoherent to prescribe it as a goal. The freedom from all goals and purposes is not "up to us" in the sense that it would be practically inconsistent for us to strive for it as if it were a goal or a purpose, not because our future is determined by an ineluctable fatality. Hence the conflict between the ethical and the scientific versions of the eternal recurrence remains unresolved.

Löwith's proposal presents other serious problems. For instance, he explains nowhere why appreciating the significance of the eternal recurrence as a scientific theory would lead us to a *love* of our fate rather than to a resigned acceptance of it. Recognizing the ineluctability of our fate might well induce resignation, and resignation is certainly an acceptance of our fate. The question is whether it also amounts to a "love" of it and this is simply not so clear; for we may well resign ourselves to a life we do not love. Indeed, the very idea of resignation connotes surrender to an existence we would have preferred to be different. It is also, in Nietzsche's eyes, a paradigmatic instance of the "negation of life."

Finally, the book's argument is founded on a fundamental but unsupported assumption. As an "ethical imperative," the doctrine of the eternal recurrence presupposes that we are agents responsive to imperatives, a fact the "scientific theory" of the eternal recurrence denies. The two aspects of the doctrine are in conflict, however, only if we assume that the question of whether or not we are justified to regard ourselves as agents is a *metaphysical* question that can be fully settled by an appeal to *natural facts* about us. In this case, the fact that our existence is governed by natural necessity undermines our agency. If, by contrast, we regard the question as a thoroughly *normative* question, the conflict might well be thought to disappear, for natural facts cannot settle normative questions: the fact that my life is fated to unfold in a certain way has no bearing on the question of whether I ought to consider myself an agent responsive to practical imperatives.[4] Löwith's critique of Nietzsche's doctrine thus rests on an assumption that may be challenged.

Yet Löwith merely asserts that the two versions of Nietzsche's doctrine do conflict, that he recognized this conflict, and that he undertook to resolve it. We have seen that the "solution" Löwith attributes to Nietzsche remains unsatisfactory. Furthermore, the texts that he cites in regard to the alleged conflict nearly always suggest either that Nietzsche did not recognize the conflict, or that he did not see it as a conflict in the first place (see, for example, 87). Was Nietzsche simply confused? Or did he perhaps not accept Löwith's implicit assumption on the implications of natural necessity for agency? To find a use for Löwith's critical efforts, we must tend to these questions first.

In addition to a comprehensive interpretation of Nietzsche's philosophy, the book also discusses its relations to ancient philosophy (chapter 4), to Nietzsche's own early writings (chapter 5), to modern philosophy in general (chapter 6), to comparable philosophical views in G. W. F. Hegel and Søren Kierkegaard (chapter 7), and it offers, in an extensive Appendix, a history of the interpretation of Nietzsche from 1894 to 1954—all of which the specialist might find useful.

---

[4] This argument is made notoriously by Peter Strawson in "Freedom and Resentment," *Proceedings of the British Academy*, XCVIII (1962): 1-25.

My outline of the philosophical core of Lōwith's book leaves out a number of observations and suggestions which will intrigue the student of Nietzsche and which may even provide ways of solving some of the difficulties I have discussed. But they are presented in a manner likely to frustrate, by its lack of clarity and philosophical detail, readers used to the prevailing standards of philosophical writing. Nevertheless, Lōwith's book will inspire in the persevering reader a new appreciation of some of the central issues in Nietzsche's philosophy—which is, presumably, the distinctive merit of a "classic."

BERNARD REGINSTER

Brown University

NEW BOOKS: TRANSLATIONS

ADORNO, THEODOR W.: *Critical Models: Interventions and Catchwords.* HENRY W. PICKFORD, trans. New York: Columbia, 1998. xii, 404 p. Cloth $37.50.

ARISTOTLE: *Politics.* C.D.C. REEVE, trans. Indianapolis: Hackett, 1998. lxxix, 288 p. Cloth $34.95, paper $8.95.

BLAY, MICHEL: *Reasoning with the Infinite: From the Closed World to the Mathematical Universe.* M.B. DEBEVOISE, trans. Chicago: University Press, 1998. 216 p. Cloth $30.00.

CONFUCIUS: *The Original Analects: Sayings of Confucius and His Successors.* E. BRUCE BROOKS and A. TAEKO BROOKS, trans. New York: Columbia, 1998. x, 342 p. Cloth $29.50.

DILWORTH, DAVID A., VALDO H. VIGLIELMO, and AGUSTIN JACINTO ZAVALA, eds. and trans.: *Sourcebook for Modern Japanese Philosophy: Selected Documents.* Westport, CT: Greenwood, 1998. xx, 428 p. Cloth $95.00.

GHĀZĀLĪ, AL-: *The Incoherence of the Philosophers.* MICHAEL E. MARMURA, trans. Provo, UT: Brigham Young, 1997. xxxi, 260 p. Cloth $29.95.

HEIDEGGER, MARTIN: *Pathmarks.* WILLIAM MCNEILL, ed. New York: Cambridge, 1998. xiii, 385 p. Cloth $59.95, paper $19.95; *Phenomenological Interpretation of Kant's* Critique of Pure Reason. PARVIS EMAD and KENNETH MALY, trans. Bloomington: Indiana UP, 1997. xxi, 296 p. Cloth $35.00.

HENAFF, MARCEL: *Claude Levi-Strauss and the Making of Structural Anthropology.* MARY BAKER, trans. Minneapolis: Minnesota UP, 1998. 292 p. Cloth $49.95, paper $19.95.

KANT, IMMANUEL: *Groundwork of the Metaphysics of Morals.* MARY GREGOR, ed. and trans. New York: Cambridge, 1998. xxxvi, 76 p. Cloth $34.95, paper $12.95.

KIERKEGAARD, SØREN: *The Book on Adler.* HOWARD V. HONG and EDNA H. HONG, eds. and trans. Princeton: University Press, 1998. xix, 403 p. Cloth $65.00; *The Point of View.* HOWARD V. HONG and EDNA H. HONG, eds. and trans. Princeton: University Press, 1998. xxvii, 351 p. Cloth $55.00; *The Moment and Late Writings.* HOWARD V. HONG and EDNA H. HONG, eds. and trans. Princeton: University Press, 1998. xxxi, 678 p. Cloth $65.00; *Prefaces: Writing Sampler.* TODD W. NICHOL, ed. and trans. Princeton: University Press, 1998. xviii, 211 p. Cloth $39.50.

KLOSSOWSKI, PIERRE: *Nietzsche and the Vicious Circle.* DANIEL SMITH, trans Chicago: University Press, 1998. xx, 282 p. Cloth $50.00, paper $19.95.

LACOQUE, ANDRE, and PAUL RICOEUR: *Thinking Biblically: Exegetical and Hermeneutical Studies.* DAVID PELLAUER, trans. Chicago: University Press, 1998. 441 p. Cloth $30.00, paper $23.95.

LEVINAS, EMMANUEL: *Of God Who Comes to Mind.* BETTINA BERGO, trans. Stanford: University Press, 1998. xv, 211 p. Cloth $45.00, paper $17.95.

LONGUENESSE, BÉATRICE: *Kant and the Capacity to Judge: Sensibility and Discursivity in the Transcendental Analytic of the* Critique of Pure Reason. CHARLES T. WOLFE, trans. Princeton: University Press, 1998. xv, 420 p. Cloth $59.50.

NANCY, JEAN-LUC: *The Sense of the World.* JEFFREY S. LIBRETT, trans. Minneapolis: Minnesota UP, 1997. xxvi, 207 p. Cloth $49.95, paper $19.95.

THE JOURNAL OF PHILOSOPHY

VOLUME XCV, NO. 12, DECEMBER 1998

CANTORIAN ABSTRACTION: A RECONSTRUCTION AND DEFENSE*

In section 1 of his *Beiträge*, Georg Cantor[1] explains the notion of cardinal number:

> We will call by the name "power" or "cardinal number" of M the general concept which, by means of our active faculty of thought, arises from the aggregate M when we make abstraction of the nature of its various elements *m* and of the order in which they are given.
>
> We denote the result of this double act of abstraction, the cardinal number or power of M, by
>
> (3) $\overline{\overline{M}}$.
>
> Since every single element *m*, if we abstract from its nature, becomes a "unit", the cardinal number $\overline{\overline{M}}$ is a definite aggregate composed of units, and this number has existence in our mind as an intellectual image or projection of the given aggregate M (*ibid.*, p. 86).

Later in section 7 of the same work, he gives an analogous account of order type:

> Every ordered aggregate M has a definite "order type", or more shortly a definite "type", which we will denote by
>
> (2) $\overline{M}$.

---

* I should like to thank the members of a philosophy of mathematics seminar at University of California, Los Angeles, and of a methodology seminar at Columbia University for their many helpful comments. I am especially grateful to Tony Martin and David Kaplan

[1] "Beiträge zur Begrundung der transfiniten Mengen-lehren," translated as *Contributions to the Founding of the Theory of Transfinite Numbers*, Philip E. B. Jourdain, trans. (New York: Dover, 1955).

0022-362X/98/9512/599–634

> By this we understand the general concept which results from M if we only abstract from the nature of the elements m, and retain the order of precedence among them. Thus the order type $\overline{M}$ is itself an ordered aggregate whose elements are units which have the same order of precedence amongst one another as the corresponding elements of M, from which they are derived by abstraction (*ibid.*, p. 111–12).

Richard Dedekind, in section 73 of his *Was sind und was sollen die Zahlen*,[2] gives a somewhat similar account of number series:

> If in the consideration of a simply infinite system N set in order by a transformation ϕ we entirely neglect the special character of the elements, simply retaining their distinguishability and taking into account only the relations to one another in which they are placed by the order-setting transformation ϕ, then are these elements called ***natural numbers*** or ***ordinal numbers*** or simply ***numbers***, and the base-element 1 is called the ***base-number*** of the ***number-series*** N. With reference to this freeing the elements from every other content (abstraction) we are justified in calling numbers a free creation of the human mind (*ibid.*, p. 68).

What these accounts have in common is a view of abstraction as the process of freeing an object of its peculiar features and a conception of number or order type as the product of such a process. Natural as this approach to abstraction may be, it has not met with much favor among contemporary philosophers. Gottlob Frege subjected it to scathing criticism in sections 28-44 of the *Grundlagen*[3]; and most philosophers today would agree with Michael Dummett's[4] view of the theory as "misbegotten" and his assessment of Frege's critique as "conclusive" (*ibid.*, p. 80).

I know of only two attempts at defense in the recent philosophical literature. Michael Hallett[5] finds in Cantor an anticipation of J. von Neumann's approach to cardinals as sets of ordinals. But by this he means no more than that "to make sense of Cantor's abstractionism as a *set* theory of number, one is forced to bring in strong ordinal assumptions" (*ibid.*, p. 1) (a claim that we shall later see is mistaken). William Tait,[6] if I understand him correctly, sees Dedekind and Cantor as providing a "reductive" account of number: to talk of numbers in the one case is to talk of the elements of a given omega progression as if they had no properties beyond those of successor; and to talk of cardinal numbers in the other case is to talk as if there were no difference between equinumerous sets. Such an interpretation will perhaps fit Dedekind's words but, as Tait himself concedes (*ibid.*, p. 96), it makes Cantor's conception of cardinal numbers as sets of pure units both ill-motivated and ill-conceived.

---

[2] *Essays on the Theory of Numbers, I: Continuity of Irrational Numbers; II: The Nature and Meaning of Numbers*, W. W. Beman, trans. (La Salle, IL: Open Court, 1948).

[3] *The Foundations of Arithmetic: A Logico-Mathematical Enquiry into the Concept of Number*, J. L. Austin, trans. (Evanston, IL: Northwestern, 1968, 2d ed.).

[4] *Frege: Philosophy of Mathematics* (Cambridge: Harvard, 1991)

[5] *Cantorian Set Theory and Limitation of Size* (New York: Oxford, 1984).

[6] In *Frege: Importance and Legacy*, M. Schirn, ed. (Berlin: de Gruyter, 1996), pp. 93-96.

The fault of these interpretations lies, to my mind, in their failure to recognize the distinctive ontology that lies behind these accounts. Abstraction, as Cantor and Dedekind conceive it, is ontologically innovative—it leads to objects that are genuinely new. But Hallett and Tait are presumably so reluctant to accept such an ontology themselves that they do their best not to attribute it to others: Hallett by seeing the apparently new objects as objects of a familiar sort, and Tait by treating the apparent reference to them as a mere façon de parler. As I hope will become clear, however, it is only by taking the ontology seriously that we can come to a satisfactory view of what these authors had in mind.

Moreover, once this is done, it will emerge that Cantor's and Dedekind's accounts of number and order type are as coherent, interesting, and plausible as the more familiar accounts of Bertrand Russell and Frege and of von Neumann and Ernst Zermelo. Whether they are correct is, of course, another matter. The defense provided in this paper is merely of their coherence and relative plausibility, not of their correctness. But I should state for the record that I am inclined to accept the Cantorian account of order type though not the Cantorian or Dedekindian account of number.

In what follows, I shall concentrate on the views of Cantor, though it should be clear how what I say can be modifed to apply to those of Dedekind. I have not attempted to capture all of the nuances or tensions in Cantor's thought but merely to develop what I take to be its spirit, or central idea. And in developing this idea, I have been guided more by what the idea itself requires than by Cantor's own writings.

The plan of the paper is as follows. I begin by setting out what appear to be decisive objections to the Cantorian account. I then show how these objections can be overcome by making use of the theory of arbitrary objects developed in my *Reasoning with Arbitrary Objects*.[7] The relevant parts of the theory are outlined in section II; and the application to Cantor's account of number is made in section III. I show, in section IV, how the approach may be extended to order types and to structure types in general. In the final two sections, I first compare the Cantorian approach to abstraction with the standard approaches of von Neumann and Zermelo, on the one side, and of Russell and Frege, on the other; and I then consider to what extent the Cantorian approach is capable of yielding a structuralist conception of number and order type. In a formal addendum, I briefly indicate how the present theory might be formalized within an extension of Zermelo-Fraenkel set theory.[8]

---

[7] Cambridge: Blackwell, 1985

The present line of investigation ties in with certain other topics that I shall mention but not explore. First, the theory of units developed in the addendum corresponds to a theory of classes in which classes can naturally be taken to be members of sets. This theory is, I believe, of considerable interest in its own right. Second, the Cantorian theory can be extended to provide a more general theory of types—covering not merely the abstract formal types of mathematics but also the more concrete types of ordinary and scientific discourse. Third, the theory of arbitrary objects is the only intuitive theory that I know of to yield an *arbitrarily* large number of indiscernible objects. This possibility is of obvious philosophical interest and can be used to provide a justification for Harvey Friedman's[9] "complete theory of everything." Fourth, the present theory may provide a rigorous basis for understanding certain ancient views on 'one' and number. I hope to write elsewhere on some of these other topics.

I. OBJECTIONS TO CANTORIAN ABSTRACTION

In attempting to reconstruct Cantor's account of cardinal number, I shall make a certain simplification. Cantor supposes that sets are given with a certain order on their elements. He therefore takes the cardinal number $\overline{\overline{M}}$ of an aggregate M to be the result of a *double* abstraction on M (hence the double bar). One first abstracts from the nature of the elements in M to obtain its order type $\overline{M}$, and then abstracts from the order in $\overline{M}$ to obtain the cardinal $\overline{\overline{M}}$. I shall suppose, on the other hand, that we start off with an unordered set M and obtain the cardinal number of M as the result of a *single* act of abstraction on its elements.

It might be thought that, by proceeding in this way, we deprive ourselves of a significant advantage of Cantor's own account. For if the cardinal number $\overline{\overline{M}}$ of M is derived from an order type $\overline{M}$, then the position of the units in $\overline{M}$ may give us a clue to their identity and hence to the identity of the units in $\overline{M}$. Now, this point might indeed have some merit if the given orderings were well-orderings or if, in some other way, they assigned a unique position to each of their elements; for we might then identify the unit in terms of its position within the ordering. But the orderings, for Cantor, comprise all linear orderings—whether well-ordered or not; and so if M were taken to be the set of rationals under their natural ordering, then it is hard to see how M could be of any more help in fixing the identity of the units than the corresponding unordered set of rationals.

---

[8] I use the following notation throughout the paper (addendum apart). The letters x, y, z,... are reserved for variable objects, the script letters x, y, z,...for variable signs, and a, b, c,... for objects in general. The value of a variable object may be indicated by underlining the sign for the variable or subscripting it with '0'.

[9] "A Complete Theory of Everything: Validity in the Universal Domain," at www.math.ohio-state.edu/~friedman/.

I am therefore inclined to think that, if Cantor's account can be made to work, then so can an account that takes unordered sets as its starting point. Moreover, the alternative account is superior to Cantor's on at least two counts. First, given the notion of cardinal number for an unordered set, we can define the cardinal of an ordered set as the cardinal of the corresponding unordered set (without having to go through the order type). On the other hand, given the notion of cardinal number for an ordered set, one can only define the cardinal of an unordered set (such as a cardinal itself, on Cantor's view) if one assumes that there *is* a corresponding ordered set. But that every set can be ordered is a contentious assumption which it would be better to do without. Second, unordered and ordered sets are special kinds of structures. One would like to have a general acccount of structure type and it is hard to see how such an account could succeed unless it *directly* yielded the type of each kind of structure.

There are several significant objections that can be made to Cantor's account, whether stated in terms of single or double abstraction. Some of them are foreshadowed in Frege's *Grundlagen*, but I have not attempted to determine the exact relationship between my own objections and his. The first, and most glaring to contemporary philosophical sensibility, is that the account is stated in psychological terms. The units are portrayed as the product of a psychological act of abstraction; *we* abstract from the elements of a set; and the resulting number has some kind of existence in our mind "as an intellectual image or projection" of the set. The notion of abstraction would therefore appear to be a psychological notion. But if we are to have a definition of number, then it should be stated in mathematical, or logico-mathematical, terms and certainly not in psychological terms.

Cantor himself had previously acknowledged the need to keep psychological elements out of the concepts and propositions of arithmetic. In reviewing Frege's *Grundlagen*,[10] he writes:

---

[10] Cantor, Untitled Review, *Deutsche Litteraturzeitung*, xx (May 16, 1885). 728-29.

> It should also be recognized that the author has adopted the correct position in so far as he has demanded that spatial as well as temporal intuition and likewise all psychological elements must be kept out of the concepts and basic propositions of arithmetic. Only in this way can rigorous logical purity be achieved, and thereby the justification for the application of arithmetic to imaginable objects of cognition (*ibid.*, p. 440).

So perhaps Cantor was not attempting to provide a *definition* of number in the passage quoted earlier. Perhaps for him the notion 'x is the cardinal number of y' is primitive and what we have is not a definition of the notion but a psychological account of how it is to be grasped.

This interpretation does not seem to square with Cantor's own understanding of the passage, however; for he treats it as providing a "definition of power" and appeals to it as such in his proof that equinumerous sets are of the same cardinality (*op. cit.*, p. 83). This therefore suggests that the psychological detail was incidental to his purposes; and, indeed, in the parallel definition of order type in part two of the *Beiträge*, the psychological element is barely present.

Whatever Cantor's view might have been, the value of his proposal will rest on whether a properly mathematical notion of abstraction can be made out and whether number can be defined, or otherwise explained, in terms of such a notion. Unfortunately, Cantor's own formulations are not of much help in this regard. For once they are stripped of their psychological trappings, nothing much seems to remain; the apparent substance of the account appears to evaporate in mental mist.

Indeed, it may be argued that no notion of Cantorian abstraction can properly meet the demands imposed upon it and that any attempt to explain number or order type in terms of abstraction must therefore be in vain. Let us consider two objections of this sort, one focusing on the notion of abstraction itself and the other on the correlative notion of 'unit'.

The first objection amounts to a proof that there can only be two numbers under the Cantorian account: 0 and 1. For let $\alpha(a)$ be the result of abstracting on the object a. The number of any set $\{a_1, a_2,...\}$ is then $\{\alpha(a_1), \alpha(a_2),...\}$. Consider now any two objects, say Socrates and Plato. Then the number of {Socrates} is {α(Socrates)} and the number of {Plato} is {α(Plato)} and, since these two numbers are the same, α(Socrates) = α(Plato). But then the number of {Socrates, Plato} is {α(Socrates), α(Plato)}, which is the same as the number of {Socrates}! As Frege saw so clearly in the *Grundlagen* (§45), the account of number in terms of units ascribes to units two contradictory qualities, namely, identity and distinguishability.

One might attempt to overcome this problem by making the abstraction function relative to the cardinality of the set in question. Thus, how one abstracts will depend upon how many objects one is abstracting from. But this, too, will not do. For let N be the set of nonnegative integers and P be the set of positive integers. Then the cardinality c of the two sets is the same and yet the number $\{\alpha_c(1), \alpha_c(2),...\}$ of P will be a proper subset of the number $\{\alpha_c(0), \alpha_c(1),...\}$ of N.

It will not even do to make the abstraction function relative to the *set* in question. For consider the doubleton X = {Socrates, Plato}. Its cardinal number is the set of units $\{u_0, u_1\}$. But on what basis do we set α_X(Socrates) identical to u_0 rather than u_1 and, given that the cardinal number of Y = {Caesar, Brutus} is also $\{u_0, u_1\}$, then on what basis do we set α_Y(Caesar) identical to α_X(Socrates) rather than to α_X(Plato)?

The argument may be stated in a more positive fashion. If we have properly abstracted from the members of a set, then the resulting units should be indistinguishable from one another—for abstraction should eliminate any distinctions among the objects to be abstracted beyond their merely being distinct. If, for example, the units u_0 and u_1 are the result of abstracting on the objects a_0 and a_1, then no distinction between a_0 and a_1, beyond their merely being distinct, should persist as a difference between u_0 and u_1. Cantor[11] himself seems to endorse such a view when he writes:

> the elements [of the cardinal number] as so-called ones have in a certain way grown together organically into one another to a *uniform whole* [italics added] in such a way that none of them has a privileged relation of rank to the others (*ibid.*, p. 379).

But abstraction, even when it is made relative to the set upon which the abstraction is performed, violates this requirement of indistinguishability. For consider the number $\{\alpha_X$(Socrates), α_X(Plato)$\}$ that results from abstracting on the members of the set X = {Socrates, Plato}. If α_X(Plato) were distinct from α_X(Socrates), then it would be distinguished from α_X(Socrates) by virtue of the fact that it was the result of abstracting on Plato rather than on Socrates. Thus indistinguishability without identity appears to be impossible.

---

[11] "Mitteilungen zur Lehre vom Transfiniten I," *Zeitschrift für Philosophie under philosophische Kritik*, XCI (1887): 81-125.

To remedy this further difficulty, one might think that abstraction should be taken to apply to the elements of a set as a whole. Thus, rather than regard the number of a set $\{a_1, a_2,....\}$ as the result of the *multiple* applications of the abstraction function to the individual elements $a_1, a_2,...$, we now regard the number as the result of a *single* application of the function to all of the elements $a_1, a_2,...$ at once. There is even some evidence that Cantor himself may have favored such an approach; for in reference to Leibniz's comments on number, he writes:

> The addition of ones, however, can never serve for a definition of a number, since here the specification of the main thing, namely *how often* the ones must be added, cannot be achieved without using the number itself. This proves that the number is to be explained only as an organic unity of ones achieved by a *single act of abstraction* [italics added] (*ibid.*, p. 381, n. 1).[11]

But how are we to construe this single act of abstraction? Presumably, in order to give general expression to the idea that certain objects are the abstractions of certain others, taken as a whole, we must take the *set* of the former to stand in the appropriate relation to the *set* of the latter. But this relation between sets *is* just that of a cardinal number to the sets that are of that cardinal number. It is therefore no longer clear how the one notion might be used to explain the other.

The second objection I wish to consider turns on the internal constitution of the numbers themselves rather than on the external relationship of the units to other objects. The Cantorian number 1 is a set $\{u\}$ consisting of a single unit u. The Cantorian number 2 is a set $\{v, w\}$ consisting of two units v and w. Since the units v and w are two in number, at least one of them, say v, is not the unit u. But why should 1 consist of the unit u rather than the unit v? If we abstract on $\{u\}$, we obtain $\{u\}$. But if we abstract on $\{v\}$, we obtain $\{u\}$. Why should this be? Why should certain units be especially suited to compose a number or to be the products of abstraction and not others?

One might think that there was an easy answer to this question; for we may suppose that the units are generated in turn in much the same way as the natural numbers. Thus there will be a first unit u_1, its successor u_2, *its* successor u_3, and so on. This then makes it natural to suppose that when we abstract from objects we go through the units in turn. The resulting natural numbers will therefore consist of sequences of units: 1 will be $\{u_1\}$; 2 will be $\{u_1, u_2\}$; and so on. We thereby obtain a von Neumann style version of the Cantorian theory. But there is no need to identify u_n with the number n-1 as on von Neumann's own account. We could instead imagine that u_1 was an "original" unit, u_2 a "copy" of u_1, u_3 a copy of that copy, and so on.

---

[11] The passage is quoted in Hallett, p. 131, where further textual evidence bearing on the issue is discussed.

There is, however, a general objection to any abstractionist view which would make one nonzero number a subset of another or which would even allow two numbers to overlap; for recall that the units within any given number should be indistinguishable from one another. Suppose now that one nonzero number was included in or overlapped with another. Then the numbers would contain a common unit *u* and one of the numbers would contain a unit *v* not contained in the other. The units *u* and *v* of that number would then be distinguished by virtue of the fact that *u* belonged to the other number while *v* did not.

We face an additional problem in connection with order types, for which the generation of the units provides no help. For we need to explain not only why it is these units—rather than others—that constitute the type, we also need to explain why they are ordered in the way they are. Consider the order type of the rationals, for example. Why should the type be one in which the units enter into the ordering in the way they do rather than in the reverse ordering or in any of the other possible rearrangements of the given ordering?

This problem even arises for the finite ordinals (that is, the order types of finite well-orderings). Consider the finite ordinal 2, for example. It contains two units, v_1 and v_2. But why should we take one to be first and the other second, rather than the other way round? Given that one of the units belongs to the ordinal 1, it is natural to take this unit to be first. But this is merely the result of a convention; we read the ordering relation as 'less-than'. If we read it as a 'greater-than' and treated the ordinals as growing 'backward', then it would be just as natural to take the unit from the ordinal 1 to come last rather than first.

But even if this difficulty over the choice of the ordering could somehow be resolved, a difficulty would remain. For if each of the finite ordinal numbers 0, 1, 2,... is to be included in its successor and also in ω (the order type of the natural numbers under their natural ordering), then presumably each should also be included in ω^* (the order type of the nonpositive integers under their natural ordering). For how can the "direction" of the relation make a difference to whether a relation of inclusion holds? But it can now be shown[13] that no matter how we choose the units or order them within a type, it is impossible for the finite ordinals to be included, in the required way, both in their finite successors and also in ω and ω^*. This shows how special a von Neumann style of representation is to the case of the ordinals, for it breaks down once ω^* is thrown into the picture.

If we ask the general question—'When, if at all, should one type be included in or overlap with another?'—then the only possible answer would appear to be: when it or a suitable part of it can be suitably embedded in the other. But this answer cannot hold in the general case since it cannot hold in the special case above of the finite ordinals with ω and ω^*. Thus, the only reasonably systematic account of abstraction along Cantorian lines must be one that denies that inclusion or overlap of types is ever possible.

We see from the above discussion that any satisfactory account of number along Cantorian lines must solve certain problems. Most important of all, it must be shown to rest upon a cogent account of abstraction—one that is stated in mathematical (or logico-mathematical) terms, that does not already presuppose the concept of number, and that provides us with a clear and definite understanding of what abstraction is.

But there are some more particular difficulties. First, we must explain how a number can result from abstracting on the elements of a set even though no units can be uniquely associated with each element of the set. Second, we must explain how the units within a number are indistinguishable from one another even though the units from different numbers are not. Third, we must explain why only certain units belong together in a number or order type and why the units within an order type are arranged in the way they are.

Given the seemingly intractable nature of these difficulties, it is not surprising that logicians and philosophers have despaired of developing a satisfactory account of number along these lines.

II. THE THEORY OF VARIABLE OBJECTS

I believe that these difficulties may be solved by appealing to the theory of variable (or arbitrary) objects developed in my *Reasoning with Arbitrary Objects*; for the units appearing in numbers and order types may be taken to be variable objects of a suitably special kind. I shall here give an exposition of the relevant portion of the theory, though it is worth noting that the theory was originally developed with completely different applications in mind.

---

[13] The proof is as follows. The unit *u* of 1 appears in each of 1, 2,.... So either infinitely many units follow *u* or infinitely many units precede *u* in the subsequent finite ordinals. In the first case, the finite ordinals cannot all be included in ω^*; and in the second case, they cannot all be included in ω.

It is natural nowadays to think of variables as a certain kind of sign. I wish to think of them, however, as a certain kind of object. Like variable signs, these objects will take other objects as values; but unlike variable signs, they will not be linguistic in character. They will not be conventional symbols for their values, but abstract objects which assume those values by way of their intrinsic nature.

This difference in the ontological status of variables will be critical for our purposes. For once it is allowed that variables are objects, it will be natural to suppose that some of them are mathematical objects and hence capable of playing the role of units in a mathematical account of number. If variables were signs, they would be entirely unsuited for this purpose. We would have a form of "linguisticism" that was just as bad as the psychologism from which we had been trying to escape.

The variable signs of logic take values independently of one another; what value one variable sign takes is not constrained by the values that other variable signs take. I wish to allow for the possibility, however, that the value of one variable object may depend upon the value taken by others. I shall allow, for example, that there are variables x and y whose values are any real numbers x_0 and y_0 for which $x_0 = -y_0$.

There are two different ways in which we may think of one variable as depending upon the other. On the one hand, we can think of the dependence as two-way, with the values of the variables being simultaneously constrained. On other hand, we can think of it as one-way, with the values of one variable being given in terms of the value of the other. Thus we may either think of the variable x and y above as being simultaneously constrained by the requirement that the value of one should be the negative of the value of the others; or we may think of the value of one of the variables, say y, as being given as the negative of the value of the other.

A simpler account of cardinal numbers and order types can be obtained by requiring dependence to be two-way but a somewhat more satisfactory account is obtained by allowing dependence to be one-way as well. In what follows, I shall concentrate on the two-way case but also indicate how the account should proceed in the one-way case.

A (two-way) *system* of variables is a set of variables with the property that (i) any two variables in the set are co-dependent and (ii) no variable in the set co-depends upon any variable not in the set. A set of variables satisfying condition (ii) might be said to be *independent*. A system will then be an independent set of variables that contains no proper independent subset.

There is a natural criterion of identity for systems of variables. Let S and S' be any two systems of variables; and let us say that S and S' are *similar* under a one-one map $x \rightarrow x'$ from the variables of S onto the variables of S' if the values taken by the variables of S are the same as the values taken by the corresponding variables of S'. Thus, if S consists of the variables x_1, x_2,... and S' of x_1', x_2',..., then the systems will be similar if, for any objects a_1, a_2,..., a_1, a_2,... will be the respective values of x_1, x_2,... just in case they are the respective values of x_1', x_2',.... The criterion of identity then states that the variables of similar systems are the same. The identity of a system of variables is given by the values that the variables assume.

Consider the special case in which S contains the single variable x and S' the single variable x'. Thus x and x' will be independent variables, depending upon nothing but themselves. The criterion then says that the variables x and x' are the same as long as their values are the same.

Consider now the special case in which S contains the two variables x and y while S' contains the two variables x' and y'. The criterion then says that the variables $\{x, y\}$ will be the same as the variables $\{x', y'\}$ as long as the objects a, b are values of x, y just in case they are values of x', y'.

The criterion does not state that, in this case, $x = x'$ and $y = y'$. Indeed, consider the case above in which the value of y is the negative of the value of x. Then a, b will be a value of x, y just in case it is a value of y, x. But we do not want to say that $x = y$.

In the case that one-way dependence is allowed, the criterion for the identity of two systems of variables S and S' is somewhat more complicated. In the first place, we should require that the variables in S and S' unilaterally depend upon the same variables. And in the second place, we should require that, under some one-one map between the variables of S and S', the systems should be similar relative to any given values of the variables upon which they depend. Thus, if x and y each unilaterally depend upon a single variable, then they will be the same as long as that variable is the same and, for each value of that variable, their values are the same.

We should note that the above criteria are false for variable signs. Two independent variable signs, for example, may take the same values and yet be distinct. But even though the signs are distinct, we take the corresponding variable objects to be the same. This is another respect in which it is critical to our purposes that variables should be objects rather than signs.

I have so far talked about the abstract theory of variable objects without reference to a language by which they might be defined or described. There are two linguistic aspects of the theory that will be important in what follows.

First, there is a standard way in which systems of variable objects may be defined by means of conditions on their values. These conditions will contain uninterpreted variable signs to which the definition then assigns appropriate variable objects. Consider, as an example, the definition:

LET x and y be such that: x and y are reals and $y = -x$.[14]

The embedded condition will define a system of two co-dependent variable objects x and y whose values are all reals a and b for which $b = -a$. The values assumed by x and y are exactly those which satisfy the condition when assigned to x and y. It follows from the identity criterion above that the system of variable objects determined by a definition will be unique. The *assignment* of variable objects to variable signs need not be unique, however. In the above definition, for example, x and y or y and x might equally well be assigned to x and y.

The corresponding definition of a dependent (but not co-dependent) negative variable is given by following two clauses:

LET x be such that: x is real
LET y be such that: $y = -x$

Here, the first clause defines the independent real, while the second clause defines the dependent real whose values is always the negative of the value of the first real. Thus, in this case, each clause picks out a unique variable object.

The second linguistic aspect concerns the use of variable signs to which a variable object has already been assigned. There are two ways to interpret sentences containing such signs. The first, which I call *representational* or *generic*, corresponds to the usual interpretation of variable signs. Whether the sentence is true will depend upon what values are assigned to the variable objects (and hence indirectly to the signs themselves). In particular, we may follow the convention that the assertion of a sentence under a generic interpretation is to be true if it is true for all values of its variables.

---

[14] The 'LET' is capitalized to distinguish this generic definition from a classical definition of the usual sort.

The second, which I call *classical* or *literal*, depends upon taking seriously the role of variable objects as intermediaries between the variable sign and its values. Under this interpretation, the sentence is taken to be true if it says something true of the variable objects signified by its variable signs. Let us suppose that the variable sign x signifies the variable object x. Then under a generic reading, the assertion:

x is a variable object

will be true only if each value of x is itself a variable object. But under the literal reading, the assertion will be true, regardless of what values are assumed by x.

III. UNITS AS VARIABLES

I shall show how one can develop a satisfactory Cantorian account of number by construing units as variables. I here present a simplified account, which will be modified in the light of the more general considerations of the following section.

The number 1 for Cantor is of the form $\{u_1^1\}$, the number 2 of the form $\{u_1^2, u_2^2\}$, the finite cardinal n of the form $\{u_1^n, u_2^n, \ldots, u_n^n\}$, and similarly for the transfinite cardinals. The problem of defining a given cardinal is the problem of saying what its units are.

Consider first the cardinal 1. It has a single member u_1^1, something that we might call *the One*. But what is the One? Intuitively, it must be something common to all objects whatever. This therefore suggests, within the framework of variable objects, that it should be a variable object whose range is unrestricted; any object whatever can be one of its values.

But this is hardly enough to single out the One, for there will be many variables whose ranges are unrestricted. Consider, for example, two variable objects which can take any two objects as values as long as those values are distinct. Then these two variables will both have unrestricted range. We might call the variable u *a unit* or a *one* if its range is unrestricted. The problem then is to say which of the ones is *the* One.

This problem may be solved by requiring the One to be an independent variable, one not depending upon any other variables. By the identity criterion for systems of codependent variables, there will be at most one unrestricted and independent variable; and so the uniqueness of the One will be assured.

The additional requirement on the One can be justified in terms of the underlying operation of abstraction. The unit u in the number $1 = \{u\}$ can be obtained by abstracting on the sole member a of any singleton set $\{a\}$. We are, therefore, in a sense, abstracting from an object that stands alone—it is, in the context of the singleton, the only object under consideration. What then corresponds to this feature of the object is that the corresponding variable u should be independent.

Consider now the number $2 = \{u_1^2, u_2^2\}$. It has two members u_1^2, u_2^2, which we may call 'the One' and *the Other*. Of course, the One in the number 2 is to be distinguished from the One in the number 1. It might therefore more accurately be called 'the One as opposed to the Other'. What then are the One and the Other?

Again, we should take each, as something common to all objects, to be a unit or variable with unrestricted range. And again, since we only abstract on two objects, we should take the corresponding variables to be independent of all other variables.

But again, these two requirements hardly suffice to determine the identity of a pair of variables v_1 and v_2. For given any particular object a, we may take the values of v_1 and v_2 to be subject only to the condition that one of them must be a. Each choice of a will then determine different pairs of variables that satisfy our two requirements.

To fix the identity of the One and the Other, we should further require that they be ***differentiated*** in the sense that the value of the One should always be distinct from the value of the Other. Thus, the objects a_1 and a_2 will be respective values of the One and the Other just in case they are distinct. The units of the cardinal 2 will therefore be an independent pair of differentiated units; and from the identity criterion it will follow that there is at most one such pair, for any one-one map between two such pairs $\{v_1, v_2\}$ and $\{w_1, w_2\}$ will yield a similarity.

This further requirement can again be justified in terms of the underlying operation of abstraction. For in abstracting from the elements of a doubleton in order to obtain the number 2, we do not wish to abstract away from all features of the two objects. We wish merely to take account of the fact that the two objects are distinct; and so it is this feature, and this feature alone, which should be preserved under abstraction.

We already see from the case of 1 and 2 the critical role played by the notion of dependence. For it is in terms of this notion that the unit of 1 and the units of 2 are to be distinguished, since the unit of 1 does not depend upon anything else, while the units of 2 depend upon one another. Without appeal to the notion of dependence, it is hard to see how we could both allow for the presence of two units in 2 and yet distinguish them from the unit in 1.

With the numbers 1 and 2 in hand, it is straightforward to extend our account to all numbers whatever. A number can simply be defined to be a system of codependent variables constrained only by the condition that their values should be distinct from one another.

For each of the finite numbers 0, 1, 2,..., it is possible to give especially simple definitions, without explicitly quantifying over variable objects themselves. 0 is, of course, the null set. Consider the definitions:

(1a) LET u^1_1 be such that: $u^1_1 = u^1_1$
(1b) Let 1 be such that: $1 = \{u^1_1\}$

The first clause defines the unit of 1 and the second clause (which is to be read classically) defines 1. Now consider the definitions:

(2a) LET u^2_1 and u^2_2 be such that: $u^2_1 \neq u^2_2$
(2b) Let 2 be such that: $2 = \{u^2_1, u^2_2\}$

The first clause defines the units of 2 and the second clause defines 2. For any natural number n, the units u^n_1, u^n_2,..., u^n_n of n and the number n itself can be defined in a similar way.

We may see from our account how the various problems that stood in the way of developing a satisfactory Cantorian theory can be solved. First, it is clear how the units within any number will be indistinguishable from one another. For the system that is the number will be similar to itself under any permutation of the units. Indeed, even if we fix the identity of all other objects in the universe, there is no way in which two units from a number system might be distinguished (except, of course, in terms of being distinct from one another). It is perhaps because of this absolute form of indistinguishability of the units within each cardinal number that there is such a strong temptation to suppose that they are all the same.

Second, it is clear how the units from different numbers will be distinguishable from one another and hence distinct; for they will be co-dependent with a different number of other units. To use phraseology suggested to me by David Kaplan, the unit of 1 will be one of one, the units of two will each be one of two, and so on for all the other numbers.

To resolve the difficulties over abstraction, we must first get clearer on what abstraction is. Suppose that we have various objects a_1, a_2,... and various features of those objects (in the form of properties and relations) which we wish to preserve under abstraction. We might represent these features by a single compendious relation R

Thus the objects $b_1, b_2,\ldots$ having the relevant features will simply consist in their standing in the relation R.

We may then say that the variables $x_1, x_2,\ldots$ result from abstracting on the objects $a_1, a_2,\ldots$ with respect to the relation R if, first, $a_1, a_2,\ldots$ stand in the relation R and, second, the variables $x_1, x_2,\ldots$ constitute a system constrained only by the requirement that their respective values should stand in the relation R. Of course, whether the variables satisfy the second condition will be independent of whether the objects satisfy the first condition. But, in general, it will only be because we already know that there are objects with the desired feature that we can be assured that there exists an abstraction with respect to that feature.

In the special case of numbers, the units $u_1, u_2,\ldots$ which comprise a given number will be obtained by abstracting on the members $a_1, a_2,\ldots$ of a set with respect to the relation of diversity. This is the relation that holds of the objects $b_1, b_2,\ldots$ just in case any two of them (in distinct positions) are distinct. Let us note that the abstraction will not result in a unique association of units with objects, since if $u_1, u_2,\ldots$ is the result of abstracting on $a_1, a_2,\ldots$ with respect to diversity then so is any permutation of $u_1, u_2,\ldots$.

Given this relational notion of abstraction, we may understand the *operation* of abstraction to be either singular or simultaneous. Suppose that the respect R with which we are to abstract is given. We may take abstraction to be an operation which, in application to each of the individual objects $a_1, a_2,\ldots$ as arguments, yields variables $x_1, x_2,\ldots$ that are the result of abstracting on $a_1, a_2,\ldots$ with respect to R. Or we may take it to be an operation which, in application to a *set* of objects $\{a_1, a_2,\ldots\}$, yields a set of variables $x_1, x_2,\ldots$ that are the result of abstracting on $a_1, a_2,\ldots$ with respect to R.

Our worry from section I, in regard to singular abstraction, was that it required us to make invidious distinctions among units. In virtue of what is the result α(Socrates) of abstracting on Socrates, say, u_1^2 rather than u_2^2? We now see that there is not a single operation of singular abstraction but as many operations as there are permutations of the individuals. Thus, in addition to the operation α, there will also be an operation β that takes Socrates into u_2^2 and Plato into u_1^2; and there will be no way to distinguish between α and β except in terms of Socrates and Plato themselves.

Our worry from section II, in regard to simultaneous abstraction, was that it would provide us with no way to understand abstraction independently of number. But our present definitions show this worry to be unfounded; for number abstraction can be taken to be a special case of abstraction in general, which can then be defined without recourse to the concept of number.

We turn finally to the question of cogency. The definitions rest, of course, upon notions from the theory of variable objects—the notions of variable, value, and dependence. These notions may not be as familiar or as established as the notions of set or membership. But my view, which I shall not defend here, is that they are already enshrined in ordinary mathematical practice and that we are capable of achieving as clear and definite an understanding of them as of the standard notions of set theory. Moreover, they surely are mathematical or logico-mathematical notions; and since the only other notions of which we made any use were from set theory and logic, our account will provide us with an explicit mathematical definition of number.

The definitions confirm not only that number is a *notion* of mathematics but also that numbers are *objects* of mathematics; for it seems plausible to suppose that a variable will be an object of mathematics if it can (in principle) be defined entirely in mathematical terms and that a set will be an object of mathematics if its members are. But since the units can be defined entirely in mathematical terms, the numbers composed of them will be mathematical objects.

IV ORDER TYPE

I consider the extension of our account to abstraction on orderings. A (simply) ordered aggregate for Cantor is what we would now call a *strict linear ordering*, one whose ordering relation is asymmetric, connected, and transitive. An ordered aggregate may be identified with an ordered pair $<A, R>$, where A is a set and R is a relation on A. We shall concentrate on the question of defining the type of the relation R; and, to this end, we shall identify R with a subset of ordered pairs from $A \times A$. Nothing will turn on the identification, however; and, once the definition is given, it will be clear how it may be extended to orderings and to other kinds of structure.

Consider, as an example, a strict linear relation with two elements. This will be of the form $\{<a, b>\}$ for a distinct from b. Its type, which we may denote by $\underset{\sim}{2}$, will also be of the form $\{<u, v>\}$, where u and v are now appropriate units obtained by abstraction from the elements a and b of a relation $\{<a, b>\}$.

But what are the units u and v (which we might designate *the first* and *the second*)? We cannot take them to be the units u_1^2 and u_2^2 of the cardinal number 2. For why should the relation type be $\{<u_1^2, u_2^2>\}$ rather than $\{<u_2^2, u_1^2>\}$? Indeed, if the relation type $\underset{\sim}{2}$ were, say, $\{<u_1^2, u_2^2>\}$, then u_1^2 would be distinguished from u_2^2 by virtue of the fact that it was the 'first' unit in $\underset{\sim}{2}$.

One might, at this point, appeal to our previous idea that the units are generated in succession. We no longer face the previous difficulty over indistinguishability, since whereas the units within the *cardinal* number 2 are not distinguishable the units within the *ordinal* 2 should be distinguishable. We may therefore take the first unit to be generated to be first in the ordering and the second to be second.

The difficulty will still arise, however, for the units within the relation type of the natural ordering for the rationals; for there is no basis upon which we can regard any unit within this ordering as first as opposed to any other. And a similar difficulty would also arise for certain types of finite relation. Consider, for example, a two-element cycle $\{<a, b>, <b, a>\}$. The units within its type should be indistinguishable, as with the finite cardinals.

Given that some of the units within a relation type are distinguishable from one another and given that the units all belong to the same system, there must be some other variable or variables within the system to which they belong; for otherwise, we would just have a system of differentiated units, that is, of units from a cardinal number, all of which would be indistinguishable from one another. And given that there are other variables in the system, it must be by reference to them that the units capable of being distinguished are indeed distinguished.

But what might this other variable or these other variables be? An important clue is provided by the intuition that the units within a relation type are not "bare" units, merely indicating one object as opposed to another, but "positional" units, which also indicate the position of an object within the relation. This suggests that the variables within a relation type should be taken to depend, not merely upon one another, but upon a relation variable whose values serve to indicate the position of the units. So in the case of the relation type $\underset{\sim}{2}$, we shall have a system of *three* variables—two units v_1 and v_2 and a positional variable R—whose respective values $\underline{v}_1$, $\underline{v}_2$ and $\underline{R}$ are subject to the constraint that $\underline{v}_1$ should be distinct from $\underline{v}_2$ and $\underline{R}$ should be identical to $\{<\underline{v}_1, \underline{v}_2>\}$. Or, again, in the case of the order type of the rationals, we will have a system consisting of a variable w_p for each rational p and a relation variable R, whose respective values—$\underline{w}_p$ for each rational p, and $\underline{R}$—should be subject to the constraint that $\underline{w}_r \neq \underline{w}_s$ for $r \neq s$ and $\underline{R} = \{<\underline{w}_r, \underline{w}_s>: r < s\}$.

There are two views one might take on the relation of dependence between the units and the relation variable. On the one hand, one might suppose that the relation variable is an independent variable, whose values are all of the relations of the given type, and that the units are unilaterally dependent upon the relation variable. On the other hand, one might suppose that the relation variable and the units are codependent upon one another. Thus, in the first case, the values of the units are given in terms of the value of the relation variable; in the second case, the values of the relation variable and of the units are given together.

It still remains to define the relation type R^+. It is not clear that we can take this to be the relation variable R. For although R is generically of the right form, it is not clear that it is literally of the right form, that is, a set of ordered pairs of the given type. We can, however, take the relation type R^+ to be the set of ordered pairs of units in which their actual position is the same as their generic position. The relation type $\underset{\sim}{2}$, for example, will be $\{<v_1, v_2>\}$ (that is, the relation in which the first precedes the second), while the type of the natural ordering on the rationals will be $\{<w_r, w_s>: r < s\}$.

As before, by using let-clauses, we may give simple definitions of the relation types in particular cases. The definition of $\underset{\sim}{2}$, for example, is given by:

(3a) LET $\mathscr{R}_2$, v_1 and v_2 be such that: $v_1 \neq v_2$ and $\mathscr{R}_2 = \{<v_1, v_2>\}$
(3b) Let $\underset{\sim}{2}$ be such that: $\underset{\sim}{2} = \{<v_1, v_2>\}$

and, more generally, the definition of $\underset{\sim}{n}$ takes the form:

(4a) LET $\mathscr{R}_n$, $v_1, \dots , v_n$ be such that: $v_1, \dots , v_n$ are pairwise distinct and $\mathscr{R}_n = \{<v_1, v_2>, \dots , <v_{n-1}, v_n>\}$
(4b) Let $\underset{\sim}{n}$ be such that: $\underset{\sim}{n} = \{<v_1, v_2>, \dots , <v_{n-1}, v_n>\}$

If the units are made to depend unilaterally upon the relation variable, the definition of $\underset{\sim}{2}$ takes the following form:

(5a) LET $\mathscr{R}_2$ be such that: $\mathscr{R}_2 = \{<u, v>\}$ for some distinct u and v;
(5b) LET v_1 and v_2 be such that: $\mathscr{R}_2 = \{<v_1, v_2>\}$;
(5c) Let $\underset{\sim}{2}$ be such that: $\underset{\sim}{2} = \{<v_1, v_2>\}$.

And similarly for the other cases.

The definition of the relation type R^+ requires the definition of the auxiliary item R, which we may call its *prototype*. The objects R and R^+ are very closely related and it is important to be clear on how they are like and unlike. R is a relation *variable*; it takes all relations of a certain sort as values. R^+, on the other hand, is a variable *relation*; it is a specific relation between variables of the given sort and hence itself a value of the relation variable R. Although R is defined as taking certain relations as values while R^+ is defined to *be* one of those values, there is a sense in which R^+ itself represents the very same values as R. Consider, by way of illustration, the relation type $\underset{\sim}{2} = \{<v_1, v_2>\}$ and let R_2 be the corresponding relation variable. Then for any given values $\underline{v}_1$ and $\underline{v}_2$ of v_1 and v_2, $\underset{\sim}{2}$ will represent the value $\{<\underline{v}_1, \underline{v}_2>\}$, since this is what is obtained by substituting $\underline{v}_1$ and $\underline{v}_2$ for v_1 and v_2 in $\underset{\sim}{2}$; and this value will be exactly the same as the value assumed by the relation variable R_2 when v_1 and v_2 take the given values v_1 and v_2.

In case the prototype R is defined to be *co*dependent with the units, its behavior will closely mimic that of R^+. R^+ literally holds between certain units. So, for example, $\underset{\sim}{2}$ relates v_1 to v_2 in the set-theoretic sense that the ordered pair $<v_1, v_2>$ is one of its members. But although R is not defined as a relation, it will hold *generically* between certain units. Thus, R_2 will hold between v_1 and v_2 in the sense that, for all values $\underline{R}_2$, $\underline{v}_1$ and $\underline{v}_2$ of R_2, v_1 and v_2, $\underline{R}_2$ holds between $\underline{v}_1$ and $\underline{v}_2$. And, in general, the condition:

(*) $\boldsymbol{v}\mathscr{R}\boldsymbol{w}$ if and only if $\boldsymbol{v}\mathscr{R}^+\boldsymbol{w}$

will be true for any variables v and w codependent with R, as long as the variable signs are interpreted generically on the left and literally on the right.

The close connection between R and R^+ under the codependency approach might suggest that the two should be identified. Thus, the relation type $\underset{\sim}{2} = \{<v_1, v_2>\}$ will be taken to be one and the same as the relation variable R that is codependent with v_1 and v_2. The identification of a type with its prototype may, in its turn, be justified on the basis of a more general principle. Suppose that S is a system of codependent variables and that each of x and y is a variable of S or is a set constructed from variables of S. The principle then says that, if x and y represent the same objects under every value assignment to the variables of the system, then x and y are the same.

I myself would not wish to endorse such a principle (I regard it with much the same suspicion as the identification of a singleton set with its member). But it is certainly possible to develop a consistent version of variable object theory within which the principle holds.[15]

---

[15] There is a difficulty though. Suppose we define u and v by:
LET $\boldsymbol{u}$ and $\boldsymbol{v}$ be such that: $\boldsymbol{u} = \boldsymbol{v}$
Then the sets $<u,v>$ and $<v,u>$ are distinct and yet always take the same values. To avoid this difficulty, we should require under the present approach that it always be possible for the variables of a system to take distinct values.

It should be clear how the above account of relation type might be generalized. Suppose that b is a set $f(a_1, a_2,...)$ built up from the distinct objects $a_1, a_2,...$ by the usual set-theoretic means. For example, b might be an ordering $<A, R>$ and the objects $a_1, a_2,...$ the elements of A; or b might be a group $<G, e, o>$ and the objects $a_1, a_2,...$ the elements of G. Any object of the same type as b (under the representation $f(a_1, a_2,...)$) will be of the form $f(z_1, z_2,...)$ for distinct objects $z_1, z_2,....$ The units and the prototype (under codependency) may then be defined by:

(#) LET $x_1, x_2,...$ and y be such that: (i) $x_1, x_2,...$ are pairwise distinct and (ii) $y = f(x_1, x_2,...)$

and the type itself by:

(##) Let y^+ be such that: $y^+ = f(x_1, x_2,...)$

Of course, when the objects $a_1, a_2,...$ are infinite in number, the definitions will also be infinitary. But there is no difficulty in converting them to a finitary form or in giving a general definition of set-theoretic type.

There are various other ways in which the account might be further extended. Condition (i) in (#), that the objects x_1, x_2 be pairwise distinct, might be dropped. We would then obtain a definition of homomorphism type rather than of isomorphism type. We might also allow the object b to be built up from the constitutent objects $a_1, a_2,...$ by other than set-theoretic means. For example, b might be the proposition that a_1 is identical to a_2. Its homomorphism type would then itself be an identity proposition to the effect that one unit was identical to another.

Indeed, all that is strictly required for definitions (#) and (##) to go through is that f should be a function that is defined upon the variables $x_1, x_2,...$ as well as upon their values. But it is not then in general possible to regard y^+ as a structural type, with a representative capacity that is independent of the function f. For this, it seems necessary that $b = f(a_1, a_2,...)$ should be a *complex* built up from constituents $a_1, a_2,....$ For such a complex there is a well-defined notion of substitution, and we may therefore take the type y^+ to represent the result of substituting the values of the variables $x_1, x_2,...$ in the type for the variables themselves. The full development of this idea would require us to graft a theory of complexes onto the theory of variables sketched below. But this is not something I shall attempt.

We should note that, if we apply the general account of structural type to sets of elements, we obtain something slightly different from our previous definition of number. The units in the number 2, for example, will now be defined by:

LET $\boldsymbol{u}_1^2$, $\boldsymbol{u}_2^2$ and $\boldsymbol{d}$ be such that: $\boldsymbol{u}_1^2 \neq \boldsymbol{u}_2^2$ and $\boldsymbol{d} = \{\boldsymbol{u}_1^2, \boldsymbol{u}_2^2\}$

Thus, instead of simply abstracting from two elements a_1 and a_2 with respect to difference, we abstract from the two elements a_1 and a_2 and their doubleton d with respect to the difference of the first two objects and their composition of the third. Since the positional variable d does not serve to distinguish between the units, it can be dropped. But this is accidental to the given case; and, in general, we will need to abstract both from the units within a given structure and from the structure itself.

We should also remark that the present approach is loosely related to the linguistically oriented approaches of Nicholas White and Hartry Field.[16] Under one variant of White's view, which brings it closer to our own, each numeral is taken to be an open term with a parameter for an arbitrary ω-progression. Thus, the numeral '0' will be construed as the term 'the first member of the ω-progression $\boldsymbol{p}$', the numeral '1' as 'the second term of the ω-progression $\boldsymbol{p}$', and so on. Under Field's approach, the numerals are taken to be a sequence of terms whose denotations are subject only to the requirement that they should always be distinct.[17] Thus, White's and Field's terms are linguistic analogues of our units. Our approach has a clear advantage, however; for it introduces an ontology of numbers; and so there is no difficulty in making sense of quantification over numbers or in referring to sets of numbers or number functions or other such higher order entities. But under the other approaches, it is not clear how the effect of such quantification or reference is to be achieved.

V COMPARISONS

The Cantorian approach enjoys certain advantages over its rivals; and I want, in this section, to lay out what these are. I begin with the technical points in its favor and then turn to more philosophical considerations. I shall argue, in particular, that the Cantorian approach is able to make a type be an instance of itself, as with the accounts of von Neumann and Zermelo, but without making an arbitrary choice of what that instance should be.

---

[16] White, "What Numbers Are," *Synthese*, XXVII (1974). 111-24; Field, "Quine and the Correspondence Theory," *Philosophical Review*, LXXXIII (1974): 200-28.

[17] Analogous treatments of instantial terms in quantificational reasoning, either as open functional terms or as "ambiguous" names, are considered in "Reasoning with Arbitrary Objects," pp. 136-42.

There are some familiar technical difficulties in von Neumann's account of cardinal numbers and in other accounts of that sort. A set will have a cardinal number only if the set can be well-ordered (since the one-one correspondence between the set and an initial ordinal will induce a well-ordering on the set). The general existence of cardinals therefore depends upon the axiom of choice or some equivalent thereof. But it would be preferable to allow for the possibility that non-well-ordered sets have a cardinal and to be able to prove, without the benefit of the axiom of choice, that every set has a cardinal.

Matters are even worse when it comes to extending the von Neumann approach to order types or to structure types in general; for it is impossible, even in the presence of the axiom of choice, to provide a suitable definition of order type. That is to say, there is no formula $\phi(x,y)$ (with the intuitive meaning 'y is the order type of x') such that:

(*) $\forall x[x$ is an ordering $\rightarrow \exists y(\phi(x,y)$ & y is isomorphic to x & $\forall x,x',y,y'[\phi(x,y)$ & $\phi(x',y')$ & x isomorphic to $x' \rightarrow y = y'])$

is provable in ZFC (proof of this claim is due to M. Foreman).

These difficulties disappear under the approaches of Russell/Frege and of Cantor. The existence of the class of sets equinumerous with a given set does not depend upon the set being well-ordered; and, likewise, nor does the existence of the corresponding set of units.[18] Moreover, the same methods that are used to prove the existence of cardinal numbers in these cases can be used to establish the existence of order types or of structure types in general.

The Frege/Russell approach cannot be accommodated within ZF, however. Indeed, the set of all sets equinumerous with a given nonempty set can be proved not to exist (since its union will be the universal set). Cardinals should therefore be classes rather than sets, and the classes in question should submit to a theory very different from ZF. But the familiar choices here—NBG (Neumann-Bernays-Gödel class theory), NF (New Foundations), or the theory of types—are not appealing. NBG will not allow cardinals to belong to sets; NF rests upon an unintuitive notion of class; the theory of types requires an ambiguous notion of cardinal; and neither NF nor the theory of types has the power or flexibility of ZF.

---

[18] It therefore appears that Hallett's remark—"With hindsight we know that the only way to obtain a representational theory [one in which each cardinal is of that cardinal] is to appeal to some strong *ordinal* notions" (*op. cit.*, p. 136)—is unwarranted.

The Cantorian approach, by contrast, can be grafted onto ZF without any difficulty; for the units can be treated as *ur*elements which can then figure in sets of the cumulative hierarchy in the usual way. Somewhat surprisingly, we see from the possibility of grafting a theory of units onto ZF that the shortcomings of the Frege/Russell approach may have been exaggerated. For there is a natural connection between variables and classes: each variable may be associated with the class of its values; and each class may be associated with an independent variable whose values are the members of that class. This therefore suggests that there should be a treatment of classes in which they are capable of belonging to sets in much the same way as variables. Indeed, using the theory of variable objects as a guide, I have been able to construct a natural and powerful theory of this sort—one which contains ZF, establishes the existence of Frege-Russell cardinals and other such types, and allows classes to belong to sets with the same freedom as *ur*elements.[19] I shall give details elsewhere, but we should remark that it was only the conception of units as a special kind of *ur*element that led to the analogous conception of classes.

Let us now turn to the philosophical considerations. Following Hallett (*op. cit.*, p. 136), we shall say that an account of the types of some kind is *representational* if each type of the given type is of that very type. The accounts of cardinal number given by von Neumann and Zermelo are representational in this sense: each cardinal number is of that number. Their accounts of ordinal number are not strictly representational, since the ordinal numbers will be sets rather than well-orderings, but they are readily rendered representational by equipping each of the sets with an appropriate well-ordering. On the other hand, the accounts of cardinal and ordinal number of Russell and Frege are not representational: each cardinal or ordinal number will not (as a rule) be of that number.

There is a certain technical advantage in adopting a representational account of numbers or of some other kind of type; for we can then reason directly with the types and define their basic properties and relations without making a detour through what they are types of. Under the von Neumann account of number, for example, the relation of less-than-or-equal-to can simply be defined as the subset relation.

---

[19] Somewhat similar theories have been proposed by Alonzo Church in "Set Theory with a Universal Set," *Proceedings of the Tarski Symposium*, Proceedings of Symposia in Pure Mathematics xxv, L. Henkin, ed. (1974): 297-308; and also by U. Oswald in "Fragmente von 'New Foundations' und Typentheorie," Ph. D. thesis, ETH Zurich (1976). But these theories are much less natural and powerful than my own. I should also mention my theory of Fregean abstraction, "The Limits of Abstraction," in *Philosophy of Mathematics Today*, Matthias Schirn, ed (New York: Oxford, 1998), pp. 503-630, which is based upon second-order logic rather than set theory and is much closer in spirit to Frege's original ideas.

There may also be a certain philosophical advantage in adopting a representational theory; for certain types do seem to be self-applicable. We do seem to conceive of the order type of a ordering, for example, as an ordering that is of that very type, though whether a cardinal number should be regarded in a similar same way as a representative item of that number is not so clear.

Along with this advantage of the accounts of von Neumann and Zermelo comes a drawback, however; for the accounts are in a certain sense arbitrary. They rest upon arbitrarily selecting certain representative items of a given number as being that number. If it is asked why should these representatives be taken to be the numbers rather than certain others—why von Neumann representatives, for example, rather than Zermelo representatives—then no satisfactory answer can be given.

One might attempt to mount a defense of von Neumann's account or ordinals as somehow being the simplest or most natural. But once we consider the general problem of representing orderings, it is hard to see how any such defense could succeed. If the von Neumann ordinals, with membership as the ordering relation, provide the simplest or most natural way of representing well-orderings, then, by considerations of symmetry, the von Neumann ordinals, with converse membership as the ordering relation, should provide the simplest or most natural way of representing the converses of well-ordering. But a finite well-ordering is of the same type as its converse; and so we must arbitrarily choose between membership and converse membership as the ordering relation on the corresponding von Neumann ordinal. To put the point differently: von Neumann identifies each element in a well-ordering with the set of its predecessors; but we can, with equal justice, identify each element in a converse well-ordering with the set of its successors.

The Russell/Frege account of numbers and ordinals as equivalence classes of sets or well-orderings does not suffer from this drawback—or, at least, not to the same extent. The account deals with the representatives in an even-handed way—each is equally taken to be a member of the equivalence class. Nor does there appear to be a really compelling argument to the effect that, if a number or ordinal is an equivalence class of the items that are of that number or ordinal, then it may, with equal justice, be taken to be a class of some other sort.

We see that the von Neumann/Zermelo style of account has the advantage over that of Frege/Russell of being representational, while the Frege/Russell style of account has the advantage over that of von Neumann/Zermelo of being nonarbitrary. The Cantorian approach, on the other hand, combines both advantages; it provides an account of number that is both representational and nonarbitrary.

It is clear that the account is representational. But whether it is nonarbitrary is not so clear. Is it not a general defect of any representational account, and not just a particular defect of the accounts of von Neumann and Zermelo, that it will involve an arbitrary choice of representatives? For representatives of the type must be chosen; and will there not always be other representatives that might be chosen with equal justice?

The difficulty might be raised in a somewhat sharper form. Any type should be what is common, in the way characteristic of a type, to all of its instances. The difficulty then is in seeing how what is common to a given range of instances can itself be one of those instances. It would then have a specificity as one of the instances that would appear to disqualify it as being what is common to all of the instances. (We have here a form of the "Third Man.")

In order to meet this challenge, we cannot simply take for granted that the items chosen to be types are what is common to their instances. We must demonstrate how this is so. And this requires that we first define (or otherwise explain) the type relation, that is, the relation between a type of a given kind and the items of which it is a type and that we then show, on the basis of the definition, how each type will, in some intuitively satisfying way, be what is common to its various instances. There is also a formal requirement that must be satisfied. For we must show, again on the basis of the definition, that each item will have a single type whose instances are all those items suitably related to the given item.

We should note that it is hard to see how one could satisfy the first two requirements, and thereby provide an intuitively satisfying account of the relation of type to instance for the types in question, without also providing a general account of the type-instance relation. Thus, one test of whether we have an intuitively satisfactory account of the relation in a given case is that it should generalize to other cases.

The three requirements can be met in the case of the Frege/Russell approach to numbers. The number of a set is the class of equinumerous sets. With 'number of' so defined, it is clear that the number of each set is unique. Moreover, with 'number of' so defined, each number will, in an intuitively satisfying way, be what is common to its instances. Indeed, the sense in which it is what is common to its instances is the same as the sense in which a class is what is common to its members; the commonality of types is but a special case of the commonality of classes. We should note that this account of 'number of' also readily yields a general account of the 'type of' relation. For given an underlying equivalence relation, the type of an item will be the equivalence class containing that item as a member.

We can now better understand how the Russell/Frege account is able to withstand the charge of arbitrariness. For effectively to challenge the idea that the number of a set is the class of equinumerous sets, we must effectively challenge the idea that a class is what is common to its members; and there appears to be no plausible way in which this might be done. If one follows Paul Benacerraf[20] and thinks of numbers as being defined by their natural ordering, then it does indeed appear to be arbitrary what classes we take the numbers to be, assuming we take them to be classes at all. But if it is built into our conception of number that they should be what is common to equinumerous sets (or equinumerous concepts or what have you), then the choice of numbers as classes of such items is not arbitrary and might even appear inevitable.[21]

The three requirements cannot, however, be met in the case of the von Neumann/Zermelo approach to numbers. For there is essentially only one way to define the 'number of' relation under this approach. One first gives an intrinsic account of number (thus on von Neumann's account, a cardinal number will be an initial transitive and $\in$-connected set); and one then defines *m* to be the number of a set M if *m is* a number and is equinumerous with M. But with 'number of' so understood, a number is not, in any intuitive sense, what is common to its instances. For given an equivalence class of numbers, how can the fact that one of its members is an initial ordinal or the like provide any basis for supposing that it is, in some intuitive sense, what is common to the other members? Indeed, we have no basis, in this case, for supposing that a number is *something* common to the other members, let alone *what* is common.

Nor is this failing peculiar to the case at hand. The accounts of von Neumann and Zermelo are of the following general sort. We are given a certain equivalence relation, let us call it *resemblance*, in terms of which the types are to be defined. Certain items from the domain of the relation are picked out as privileged in terms of their intrinsic properties (and not in terms of their relationship to the items that they resemble); and an item is then defined to be the type of another if it is privileged and resembles the other. Thus, the type-instance relation is defined as a restriction on resemblance.

---

[20] "What Numbers Could Not Be," *Philosophical Review*, LXXIV (1965): 47-73.

[21] Hale has argued for a similar conclusion, though by very dissimilar means, in section 8.II of *Abstract Objects* (Cambridge. Blackwell, 1987).

But it is hard to see how any such account could meet our requirements. For given an equivalence class of resembling items, how can the intrinsic features of one of them provide a basis for supposing that it is something common to the others?

If we are to obtain a satisfactory representational account, some other way of defining the type-instance relation must be found. It must be supposed that there is some relation which holds between a type and its instances which is other than resemblance and upon the basis of which it can be shown that the type is common to its instances; and the type-instance relation should be so defined, or so explained, that it is then clear that this other relation does indeed hold between a type and its instances. Thus, even though the type will resemble its instances, we must have an independent account of its relation to the instances, one tied to its status as a common type rather than as a resembling instance.

This is not to deny that there might also be some intrinsic way of picking out the types from among the classes of resembling items. But if this is so, then it must be true of the items picked out in this way that they stand in the required commonality-inducing relation to the items that they resemble. On a traditional conception of abstraction, for example, the color green will merely have the intrinsic property of being green—all the other intrinsic properties of green things will have been abstracted away. But if we ask 'Why should such an object be regarded as a type?', then the answer will rest on the fact that its intrinsic properties are those common to its instances.

Great as these additional demands on a representational theory may be, the Cantorian approach to types is able to satisfy them. A Cantorian type is what one might call a *form*, an abstract or nonlinguistic counterpart to a scheme. Any form can be obtained by substituting variables for the constituents in some complex, just as a scheme can be obtained by substituting schematic letters for constituents in an expression. The Cantorian number 2, for example, is the result of substituting variables for the constituents *a* and *b* in a doubleton {*a*, *b*)}; and the relation type of the natural ordering on the rationals is the result of substituting variables for the rationals in that ordering.

A form, in its turn, will have substitution instances that result from replacing the constituent variables in the form by their values. Now it is evident that the form is something common to its substitution instances, just as it is evident that a class is something common to its members. And so, given that the instances of a Cantorian type or form are the same as its substitution instances, there will be a real basis, in this case, for the claim that the type is something common to its instances.

Under the Frege/Russell approach, the underlying relation between a type and its instances was that of a class to its members and there was just one object, the equivalence class, which stood in that relation to all and only the instances of the type. But in the present case, there will be many such objects—there will be many forms that are of the given type and yet have exactly the same substitution instances as the type. For example, if *u* and *v* are two units from any number greater than 1, then {<*u*, *v*>} and {<*v*, *u*>} will both be relations of type $\underset{\sim}{2}$ whose substitution instances are exactly the relations of type $\underset{\sim}{2}$. We therefore face the problem of distinguishing that form which is the type—call it the *canonical form*—from all of the other forms and explaining why it should be regarded as *the* common element of the various instances.

When one considers the question of what the canonical form might be taken to be, there appears to be only one reasonable view. To fix our ideas, let us consider the problem of determining the canonical form of an ordered pair of distinct elements. This will itself be an ordered pair <*u*, *v*>, where *u* and *v* are variables whose values are all those distinct objects $\underset{\sim}{u}$ and $\underset{\sim}{v}$ for which $\underset{\sim}{u} \neq \underset{\sim}{v}$. Our problem, then, is to identify the variables *u* and *v*.

If the canonical form <*u*, *v*> is to be distinguished from all of the other forms <*w*, *x*>, then it must somehow be evident from the nature of *u* and *v* that they are to occupy the first and second position in the ordered pair <*u*, *v*>; and this suggests that the variables *u* and *v* must themselves in some sense be the first and second members of an ordered pair. But what can this mean? Clearly, it does not mean that they are the first and second members of *some* ordered pair, since that hardly suffices to distinguish them from any two other variables. Nor does it mean that they are the first and second members of the ordered pair <*u*, *v*>, since by same token *v* and *u* are the first and second members of the ordered pair <*v*, *u*>. All it seems can be meant is that they are the first and second members of an *arbitrary* ordered pair. Thus, it should be supposed that there is an independent variable *p*, whose values are all ordered pairs of distinct elements, and that *u* and

v are the dependent variables whose values are the first and second members of p. (In this particular case, u and v can be defined independently of one another; but in general this will not be possible.)

One might wish for a more systematic justification for the choice of the canonical form. But still, the argument we have given seems to leave little real indeterminacy in what the choice should be. Perhaps the only room for doubt is over the dependency status of the constituent variables. We have taken the variables in the type $<u, v>$ to be *dependent* upon an independent variable p whose values are all ordered pairs of distinct elements. But could we not, with equal justice, have taken these variables to be *codependent* with the variable p?

There is, however, something counterintuitive about taking the variables to be codependent (even though, for reasons of technical simplicity, we may treat them as if they were). For we think of u and v as the constituents that occupy the first and second positions in a *given* ordered pair. Thus we take what they are to be dependent upon what the given ordered pair p happens to be, and not vice versa. Moreover, this intuitive view is in line with the general policy of only assuming those dependencies required for the purposes at hand. For in order appropriately to fix the identities of u and v, we only require dependence, not codependence, upon p.

It should be noted that, under the present approach, we may provide an *intrinsic* characterization of the canonical forms and hence of the types. For suppose that f is a form $F(v_1, v_2,...)$ with variables $v_1, v_2,...$. Let v be the independent variable whose values are all substitution instances of f. Then for f to be a canonical form, we require that $v_1, v_2,...$ should be the variables whose values $\underline{v}_1, \underline{v}_2,...$, for a given value $\underline{v}$ of v, are those for which $\underline{v} = F(\underline{v}_1, \underline{v}_2,...)$. But we also have, in this case, a justification for the intrinsic characterization in terms of how the type relates to its instances. For the type will be common to its instances in the manner of a form; and the constituents of the type will be common to the constituents of the instances in the way we have described.

VI. RELATIONSHIP TO STRUCTURALISM

Let me conclude with some brief remarks on the bearing of the present approach on the issue of structuralism.[22] There are various ways to understand structuralism within the philosophy of mathematics, but the one most pertinent here concerns what one might call *representational structuralism*. Consider a categorical mathematical theory, such as the theory of numbers. Then from among all of the models that make the theory true, it is often supposed that there is a particular model—the *intended model*—which the theory is properly about. According to the representational structuralist, what distinguishes this model from all of the others is its representational role. The model somehow serves to represent all of the other models, with each element or relation of the model somehow serving to represent the elements and relations of the other models.

---

[22] Originally, this section contained a much fuller discussion but was cut at the suggestion of the editors.

Although several structuralists have been tempted by a view of this sort, they have never made clear what this representative model is meant to be. Nor have they given us any reason to believe that, from among all of the models of a categorical theory, there is indeed a unique model that is representative of the others.

It is here that our account of Cantorian abstraction may be of some help. For we may define a representational Cantorian type of a class of isomorphic models along the lines of our definition of order type. In the case of arithmetic, for example, the number 0 will be the first element in an arbitrary ω-progression, the number 1 the second element, and so on. Or, to be more exact, the prototype in this case will be an independent variable N whose values are all of the ω-progressions, the numbers 0, 1,... will be variables n_0, n_1,..., dependent upon N, whose values for any particular value $\underline{N}$ of N are the first, second,... members of $\underline{N}$, and n will be the successor of m if, for any value $\underline{N}$ of N, the value of n is the successor of the value of m in $\underline{N}$. Moreover, given such a definition, it may then be shown that the type is unique and itself a member of the class of models.

There are, however, definite limits on how far the present approach should be pushed. In the first place, even though the approach provides a defensible notion of a representative model, it does not commit one to the view that this is what any particular theory is properly about. Indeed, even if one believes that numbers are the results of Cantorian abstraction, one can still take them to be the result of abstraction on sets, as Cantor does, rather than on models. Thus, we should distinguish between *Dedekindian* abstractionism, the view that the objects of a given mathematical theory are the objects or "positions" in its representative model, and *Cantorian* abstractionism, the view that they are the result of abstracting on structurally similar objects.

Second, although this form of structuralism may be applied to a wide range of theories, it cannot sensibly be applied to the theory of variable objects itself. Under the intended interpretation of that theory, the domain will include various variables that are genuinely independent, whereas under the structuralist interpretation these, and all the other, variables in the domain will in reality be "higher order" variables whose values will depend upon what model for the theory is under consideration. It follows that any doubts one might have concerning the uniqueness of the intended interpretation for variable-object theory will transfer to the structuralist interpretation of other theories. Thus, nothing in the present approach will serve to combat radical skepticism over whether there is a unique interpretation for any given mathematical theory.

Third, the present approach will be incompatible with an austere form of structuralism, according to which the objects of the intended model possess no simple properties and enter into no simple relations beyond those given by the model itself. In the case of arithmetic, for example, the objects of the intended model will not only possess the internal property of being a number and enter into the internal relation of successor, they will also have the external property of being a variable and enter into the external relation of variable to value. Indeed, the objects of the model will be externally related by this means to every object whatever.

I myself do not find this a defect in the approach, since I doubt that there is any reasonable way in which the representational and austere forms of structuralism might be combined.

VII. FORMAL ADDENDUM

I sketch how the above application of abstraction theory may be formalized. I begin with a system that only allows for one-way dependence and then consider how to accommodate two-way dependence.

The formalization will proceed within a version of ZF with individuals. It will not much matter whether we have the axiom of choice, though we do assume foundation. There is one additional primitive 'Val', which is a one-place predicate. Intuitively, we think of Val as being true of the basic value assignments to variable objects. We often write '$v \in$ Val' though, strictly, we should write 'Val(v)'. From Val may be determined the variables and the relationships of dependence; for the variables will be the members of the domains of the basic value assignment, and two variables will be codependent when they belong to the domain of the same basic value assignment.

We use standard notions of set theory. In particular, we take a function to be a set of ordered pairs whose second member, for a given first member, is always unique; and we write 'Dom$(f) = x$' to indicate that f is a function whose domain is x. We also use the following definitions concerning variable objects.

System: 'System(S)' for 'S is the domain of some $\alpha \in$ Val'
Variable object: 'Var(v)' for 'v is in some system'
Value assignments for S: '$\alpha \in \mathrm{Val}_S$' for 'Val$(\alpha)$ & Dom$(\alpha) = S$'
Similarity of Systems under a function f: '$S \approx_f S'$' for 'System(S), System(S'), f is 1-1 function from S onto S' and, for all functions $\alpha \in \mathrm{Val}_{S'}$, $\alpha \circ f \in$ Val $\longleftrightarrow \alpha \in$ Val'

For the purpose of understanding the following definition, we should think of $[x, \phi(\alpha)]$ as a *quasi-system*, with the set x of objects in place of a system S of variables and with the "value assignments" α on x specified by the formula $\phi(\alpha)$.

Similarity of a system to a quasi-system under f: '$S \approx_f [x, \phi(\alpha)]$' for '$S$ is a system, f is a 1-1 function from S onto x and $\alpha \circ f \in$ Val $\longleftrightarrow \phi(\alpha)$ whenever Dom$(\alpha) = x$', where $\phi(\alpha)$ is a formula that does not contain S, x, or f.
Similarity of Systems: $S \approx S'$ for '$S \approx_f S'$ for some f', and '$S \approx [x, \phi(\alpha)]$' for '$S \approx_f [x, \phi(\alpha)]$ for some f'

The axioms for the system of variable objects (without one-way dependence) consist of those for ZF with *ur*elements and the following additional axioms for variables.

Functionality: $\alpha \in$ Val $\rightarrow \alpha$ is a function
Atomicity: Var$(x) \rightarrow -$Set(x)
Disjointness: $\alpha, \beta \in$ Val $\rightarrow$ [Dom$(\alpha) =$ Dom$(\beta) \vee$ Dom(α) disjoint from Dom(β)]
Identity: $S \approx S' \rightarrow S = S'$
Abstraction: $\exists\alpha(\phi(\alpha)$ & α is a function with domain $x) \rightarrow \exists S(S \approx [\phi(\alpha), x])$, where $\phi(\alpha)$ is a formula that does not contain the predicate Val.

We might think of x, in the formulation of the abstraction axiom, as a set of argument places $\{x_1, x_2, \ldots\}$. The formula $\phi(\alpha)$ then expresses a relation R whose extension consists of the tples $<\alpha(x_1), \alpha(x_2), \ldots>$ where α is any function with domain x that satisfies $\phi(\alpha)$. The system S of variables posited by the axiom are then those which can be obtained by abstracting on R.

Given our notion of function, a value assignment must always be defined on every variable in its domain. But we might broaden the notion of a function so as to allow a value assignment to be undefined on some of the variables in its domain. There are other ways in which the axiom might be strengthened. But some restriction on the form of $\phi(\alpha)$ is required; otherwise, we could prove the existence of a variable whose values are all those variables that did not take themselves as values and thereby establish a form of Russell's paradox.

We have simple definitions of cardinality. The set c will be *a cardinal* if it is a system of variables and if, for all functions α with domain c, $\alpha \in \mathrm{Val}_c$ if and only if α is one-one. The set c will then be a *cardinal of* the set x if it is a cardinal equinumerous with x.

From the abstraction axiom, it follows that every set x has a cardinal, for we may let $\phi(\alpha)$ be the condition that α is a one-one function on x. And from the identity axiom, it then follows that the cardinal of any set is unique. Similar definitions and proofs can be given for order type.

We may also give a definition of 'cardinal of' in terms of representation. Let us say that the object x is *constructed from* the system S if it is either a variable of S or an *ur*element that is not a variable of S or a set whose transitive closure contains no variables not in S. Let us use V_S for the objects that can be constructed from S. (Of course, V_S is not a set in ZF but is given by a formula; and similarly for some of the definitions below.) Now suppose that S is a system and α a basic value assignment with domain S. Then α may be extended to an assignment α^+ on V_S by means of the following rules:

(i) If x is a variable from S, then $\alpha^+(x) = \alpha(x)$

(ii) If x is a constant (that is, an urelement that is not a variable), then $\alpha(x) = x$

(iii) If x is a set, then $\alpha^+(x) = \{\alpha^+(y): y \in x\}$

Let us say that an x *represents* y if, for system S and some α with domain S, $x \in V_S$ and $\alpha^+(x) = y$. It may then be shown that each cardinal represents exactly the sets to which it is equinumerous.

In order to obtain a system that tolerates one-way dependence, we should think of the system to which a given dependent variable belongs as also containing the variables upon which it depends. The disjointness axiom now becomes:

$\alpha, \beta \in \mathrm{Val} \rightarrow [(\mathrm{Dom}(\alpha) \subsetneq \mathrm{Dom}(\beta)) \vee (\mathrm{Dom}(\beta) \subsetneq \mathrm{Dom}(\alpha)) \vee (\mathrm{Dom}(\alpha)$ disjoint from $\mathrm{Dom}(\beta))]$

In formulating the identity axiom, we require that the similarity map f between systems S and S' should be an identity on any proper subsystems. And abstraction should be supplemented with:

$\mathrm{System}(S)$ & $\forall\alpha\in\mathrm{Val}_S\exists\beta(\mathrm{Dom}(\beta) = x$ & $\beta \supsetneq \alpha$ & $\phi(\alpha,\beta)) \rightarrow \exists S^+\exists f(f$ is an identity on x & $S^+ \sim_f [x, \exists\alpha\in\mathrm{Val}_S(\beta \supsetneq \alpha$ & $\phi(\alpha,\beta))]$, where again $\phi(\alpha,\beta)$ is a formula that does not contain the predicate Val.

Consistency may be proved by the following method. We start with a standard model for set theory with *ur*elements in which the cardinality of the *ur*elements is the same as that of the universe as a whole. We pick out the definable classes of functions on any given set; and to each "similarity class" of such definable classes we then assign a system of variables from among the original *ur*elements. By using this assignment, we can define a model for the system with two-way dependence; and by appropriately iterating the procedure, we may construct a model for the system with one-way dependence.

KIT FINE

New York University

COMMENTS AND CRITICISM

ON THE PERFECTLY GENERAL NATURE OF INSTABILITY IN MEANING HOLISM*

In a recent paper, Eric Lormand[1] offers a construal of meaning holism which, if correct, would avoid the radical instability of meaning that holism seems to entail. After briefly discussing Lormand's proposal, I shall show that it fails to take seriously the perfectly general character of instability concomitant with meaning holism. As meaning holism's comprehensive instability comes to light, it becomes clear that Lormand's *multiple-meanings* approach is powerless to solve the instability problem.

I INSTABILITY AND LORMAND'S SOLUTION

Lormand correctly diagnoses the most troublesome feature of meaning holism: "If meanings are holistically dependent, then semantic differences anywhere seem to balloon into semantic differences everywhere" (51). For present purposes, let us follow Lormand in assuming a particularly strong version of meaning holism. According to it, the meaning of a representation in a system is dependent on the meanings of all the other representations in a system.[2] This would appear to make systems of representation wildly unstable: change the meaning of one representation and, it would seem, the meanings of all the other representations change, at least slightly. A further consequence is that, for any two systems, if there is any one representation meaning they do not share, then they share no meanings. If we think of systems as actual minds or speakers, holism seems to entail that no two speakers share any identical meanings if they disagree with respect to the meanings of even a single representation. The very ideas of communication and genuine agreement and disagreement look to be threatened. For ease of reference, I shall collect all of these worries about holism under the term *the basic problem*.

Lormand's significant first step in solving the basic problem is to suggest that each representation has more than one meaning. The

---

* I would like to thank Mark Warshaw and especially Steve Yalowitz for helpful comments on an earlier draft.

[1] "How to Be a Meaning Holist," this JOURNAL, XCIII, 2 (February 1996): 51-73.

[2] I am following Lormand in using the terms 'representation' and 'system'. Lormand's proposal for meaning stability and my criticism of it are unaffected if we substitute 'expression' and 'language' for 'representation' and 'system', respectively.

0022-362X/98/9512/635–40

basic problem is then solved by blocking the inference from (1) to (2), below:

(1) If a representation in a system S changes meaning, all representations in S change meaning.
(2) If a representation in S changes meaning, no representation in S has the same meaning it had before the change (57).

If we think of a token representation as having a single holistically constituted meaning, then, if any of that representation's constituents change meaning, the representation itself changes meaning. Therefore, given the holistic interdependence of representations, *no* meanings stay the same. If, on the other hand, a token representation has multiple finer-grained meanings which can be distinctly individuated, then the representation can keep many of its meanings even if one or more of its meanings change. Thus the inference to (2) is blocked when properly rewritten:

(2*) If a representation in S changes one or more of its meanings, no representation in S has any of the same meanings it had before the change.

(2*) is false on Lormand's multiple-meanings view because, again, a representation can keep *some* of its meanings even though others are revised or dropped. Finally, the proposal is relevant to the basic problem because two systems can share one or more of the meanings of a representation, thus allowing communication and genuine agreement or disagreement, without the requirement that they share *all* of the representation's meanings.[3]

What, then, is it for a representation to have more than one meaning? A natural way of thinking about a holistically constituted meaning of a representation is in terms of the beliefs or inferences associated with the representation. The meaning of my representation [cow] is determined, or constituted, by all the beliefs or inferences I associate with it. As we have seen, if my [cow] representation has only a single meaning, then, if the set of beliefs or inferences I associate with [cow] changes, so does the meaning of [cow] itself. Lormand proposes that we "think of each unit for a representation as a separable rough test for the acceptable use of that representation" (57). If we understand each test as a separate meaning of my [cow] representation, then each token representation has many meanings because it has many such tests. Sense can then be made of speakers' sharing many meanings and genuinely agreeing or disagreeing, even if they do not share an identical total set of meanings for the representation.

---

[3] Lormand notes that Akeel Bilgrami and Michael Devitt have also hit on the idea that a representation has more than one meaning (57, fn. 8). See Bilgrami, *Belief and Meaning: The Unity and Locality of Mental Content* (Cambridge: Blackwell, 1992); and Devitt, "The Methodology of Naturalistic Semantics," this JOURNAL, XCI, 10 (October 1994): 545-72. See also Devitt's *Coming to Our Senses: A Naturalistic Program for Semantic Localism* (New York: Cambridge, 1996), chapter 2, especially §2 8 Except for the following comments, I shall not explicitly take up these other versions here. Bilgrami's view is in ways more subtle than Lormand's, due to Bilgrami's thesis of the locality of content. I believe, however, that Bilgrami's theory is susceptible to the basic line of criticism presented here, though space prohibits me from extending the criticism to his views. Finally, because Devitt rejects holism, his views are not directly relevant to the present discussion.

Lormand offers the following example. Imagine there is a child named Larry whose application of his [bird] representation has the following tests: "[feathered flying animal], [thing similar enough to B1,...,B*n*] (where the [B*i*] represent alleged birds), and [thing called 'a bird' by mommy]" (58). Lormand's notion of a "test" can be understood simply as a belief. For, as Lormand says, "when Larry accepts [this is a bird] he thereby believes *that this is a feathered flying animal, that this is a thing similar enough to B1,...,Bn, that this is a thing called 'a bird' by mommy,* and so on" (58).

With these building blocks in place, we can now say how Larry can genuinely disagree, say, with his brother Moe (as opposed to using expressions with no shared meanings). Suppose that, unlike Larry, Moe does not have a [feathered flying animal] test for his [bird] representation. But like Larry, Moe does have a [thing called 'a bird' by mommy] test. If Larry and Moe disagree in their application of [bird] to an object, "they may genuinely be disagreeing about whether the object is a thing called 'a bird' by their mommy" (59). So, to whatever extent they share any tests at all for application of [bird], to that extent they share at least one meaning. This appears to solve the basic problem.

II. THE RETURN OF INSTABILITY

Notice, however, that Lormand's selection of the test that Larry and Moe share is rather misleading. It is true that, despite their potential differences in meanings associated with, say, [mommy], it is presumably fair to say that their expressions always agree in reference, namely, [mommy] refers to their mother. And despite whatever differences in their concepts associated with [called by], we all feel pretty certain that they will not effect a disagreement in their beliefs about whether their mother calls something 'a bird' if in fact she does. I grant all this; and with it I grant that Larry and Moe share at least one test if they both believe that something is a bird if it is so-called by their mother. But all this really shows is that they agree on certain instances in which to apply [bird]. The route through mommy is superfluous. Indeed, Lormand so much as says so when he considers the case of brother Curly. Curly does not share the [thing called 'a bird' by mommy] test. He appeals to his father instead. Lormand then asks what we could say on behalf of the holist when Curly shares *none* of these tests with Larry and Moe. The answer is that "we can simply add referents to this list of meanings" (58).

Clearly, Lormand is not suggesting the view that a theory of meaning requires supplementation by an independent theory of *reference*. Rather, he wants *referents* to be included in the list of fine-grained meanings. That is, without some independent theory of reference, at best what Larry, Moe, and Curly share is agreement on some instances of when to apply their respective [bird] tokens. The "misleading" part of Lormand's presentation is that he never demonstrates how Larry and his brothers might share tests for (or *beliefs about* or *inferences connected to*) [feathered flying animal] or [thing similar enough to B1,...,B*n*]. Rather, he simply *assumes* such shared tests. Of course, if, as we are assuming, meaning holism is true, then the meanings of the tests themselves are holistically constituted. But then the contents of (the beliefs that constitute) the tests will *themselves* differ if Larry and his brothers do not share identical tests for, for example, [feathered], [flying], [similar enough], and the like. That is, we have no reason to think, given holism, that the tests the brothers "share" are identical unless we have an account of how the tests for the (putatively *shared*) tests are themselves identical!

Lormand anticipates a related objection in a footnote. He replies by claiming that, with the multiple-meanings approach, holism can be reconciled with meaning atomism: "the view that the meanings of all the representations in a system can be specified completely in terms of a relatively small set of semantically atomic (or simple, primitive, basic) representations in the system" (58, fn. 10). The idea is that basic representations themselves have no tests, and so are immune to the criticism that there is no clear criterion of *shared* test. Lormand continues: "A representation can have *a* primitive meaning—and so be a semantic atom while also having *other* nonprimitive meanings—and so satisfy inferential role semantics and meaning holism" (58, fn. 10).

The meaning holist would be completely justified in not recognizing anything like her view in this proposal, regardless of its details. Indeed, this becomes even clearer as Lormand concludes the footnote by saying: "Units consisting entirely of atomic representations, then, can yield meanings which are independent of other representations in the system" (58, fn. 10). To suggest that an entire system can be specified entirely in terms of basic atoms is just to assert meaning atomism, which in turn is simply to reject holism.

One might think this conclusion too hasty. After all, Lormand's stated intention in the footnote is to reconcile two seemingly incompatible views. But in fact, Lormand's *own characterization* of the variety of meaning holism he aims to defend in the body of the paper is incompatible with his attempt at reconciliation. He writes: "I shall focus for the time being on the extreme holist view that the meanings of *all* the representations in a system are interdependent. Although I shall defend even this extreme view..., slightly less extreme positions should count as vaguely holistic" (55). If he subscribes to this view, then clearly he cannot consistently maintain the existence of "atomic representations" that are "independent of other representations in the system" (58, fn. 10).[4]

Let us return, then, to my original criticism. Our question is: In virtue of what can we say that different speakers' tests for the application of a representation are identical? The one and only response available for Lormand qua holist is to run the same multiple-meanings proposal for the tests themselves. Suppose Larry and Moe "share" tests (metalinguistically speaking) t_1 through t_9 for application of [feathered flying animal] but disagree about t_{10}, say, [thing with webbed feet]. Larry thinks (in his idiolect) that feathered flying animals have webbed feet. (Suppose he lives near a lake and most of the birds he has seen have been waterfowl.) Moe thinks (in his idiolect) that feathered flying animals have clawed feet.[5] (Suppose Moe is very young and has only seen the family parakeet and pictures of some birds in the encyclopedia.) Following Lormand, we can say that they share many meanings with respect to [feathered flying animal]. Now, suppose for the sake of argument that Larry and Moe would apply [thing with webbed feet] to the same objects when they see them, and that they both agree that a presently visible thing has webbed feet. Larry will then be prepared to apply [feathered flying animal] and [bird] to that thing. Moe will not. Therefore, even if Larry and Moe use *a* [feathered flying animal] test for their respective applications of [bird], so far we have been given no reason to think they *share* an identical [feathered flying animal] test. Indeed, it looks as though the tests are different.

---

[4] The reply that I am failing to notice the difference in kind between basic and complex representations is futile. Either both sorts of representations are meanings or they are not. If they are, then meaning holism is false because it is not the case that all meanings are interdependent. If they are not both meanings—say, basic representations are "submeanings" of some sort—then no reconciliation of meaning holism with meaning atomism is on offer. We are still left with the instability engendered by meaning holism.

[5] Notice that in saying 'metalinguistically speaking' and 'idiolect', I am trying to avoid the pitfalls into which Lormand has fallen. Perhaps Larry and Moe both use expressions like 'There is a feathered flying animal' as tests for when to assert 'There is a bird'. But the question is: What do they each *mean* by their words? (Or: What is the *content* of their representations?)

Now let us consider the case where the relations among Larry's and Moe's tests are as described in the previous paragraph, but *both* are prepared to apply their respective [bird] representations in virtue of their respective [feathered flying animal] tests being satisfied. Larry's [feathered flying animal] and hence [bird] tests apply because his [thing with webbed feet] test applies. Moe, looking at a different object, tokens his [feathered flying animal] and [bird] representations because his [thing with claws] test applies. Lormand would say that, because of their shared tests for [bird], that is, [feathered flying animal], Larry and Moe share at least one meaning of their [bird] representation. But here again, their [feathered flying animal] tests are *not* identical: they not only do not mean the same thing, they have different extensions.

The examples simply show that, unless the tests themselves are identical, Larry and Moe share tests only metalinguistically. No clear content has been given to the notion of a shared test. And if the tests themselves merely "share" some tests (however *that* is to be construed), there will be cases where the tests originally appealed to (for example, [feathered flying animal]), even if agreed upon, apply to different objects in virtue of *their* having different tests (for example, [thing with webbed feet] versus [thing with clawed feet]). Short of a proposal for shared tests, Lormand's multiple-meanings "overlapping tests" approach inherits all the difficulties of the original holistically constituted single-meaning theory. Most importantly, since the only way to share a test is to share all the tests for that test, then no two systems (/speakers) share a single test unless they share every test. A good question now is: If the instability of meaning holism infects even the more finely individuated meanings that Lormand proposes, why should the holist add to her troubles by proliferating meanings as the multiple-meanings approach counsels? The answer to this question is "she should not," which is just another way of recognizing the perfectly general nature of instability in meaning holism.

KELLY BECKER

University of California/San Diego

BOOK REVIEWS

Deconstructing the Mind. STEPHEN P. STICH. New York: Oxford University Press, 1996. viii + 222 p. Cloth $24.95.

Stephen P. Stich's three books—*From Folk Psychology to Cognitive Science*,[1] *The Fragmentation of Reason*,[2] and now *Deconstructing the Mind*—establish him as the preeminent philosophical gadfly of our time. That one of his own best-known theses is his current target, namely, the elimination of folk psychology, diminishes neither the sharpness of his arrows nor his evident pleasure in launching them. To remind us of one of his major arguments for the thesis, the book includes "Connectionism, Eliminativism, and the Future of Folk Psychology" (with William Ramsey and Joseph Garon). The other five essays are of varying lengths and publication histories. The first, longest, and title essay is virtually all new; chapters three and five—"What is Folk Psychology?" (with Ian Ravenscroft) and "Intentionality and Naturalism" (with Stephen Laurence)—appear for the first time; chapter four—"How Do Minds Understand Minds?" (with Shaun Nichols)—draws on three previous papers; the last, briefest chapter—"Naturalism, Positivism, and Pluralism"—is a revision of an earlier essay.

At risk of appearing uncharitable, I start with the last and, I think, weakest essay. Stich explains that this essay elaborates the strategy of chapter five, "Intentionality and Naturalism." Since I think it also illuminates the strategies of chapters one and three, it will be useful to get the metaphilosophical issues out before turning to the details of those essays.

In "Naturalism, Positivism, and Pluralism," Stich offers an historical thesis: the current naturalistic trend in philosophy shares with its positivist predecessor a certain type of "Puritanism." In particular, as the positivists looked for a demarcating criterion of "meaningfulness," naturalists offer a demarcating criterion of reality: "our world is a physical world and...all the things that really exist are 'nothing but' physical" (195). Starting with that assumption, he constructs an argument against naturalism and for "open-ended pluralism" (196): (1) naturalism holds that the only real entities are physical; (2) this demarcation criterion must be capable of precise formulation; (3) since there is no one relation between basic physical entities and the things we rightly believe to exist, the naturalist criterion cannot be precisely formulated; therefore (C) naturalism should be given up in favor of an open-ended pluralism, the view that "what 'legitimates' certain properties...and makes others scientifically suspect is that the former, but not the latter, are invoked in successful scientific theories" (199).

---

[1] Cambridge: MIT, 1983.

[2] Cambridge: MIT, 1990

The first two premises of this argument are as controversial as the conclusion. Why embrace premise (2)? Stich explains:

> some account must be given of the sense in which [everything is]...physical...by trying to specify a legitimating relation, R, such that a property is...naturalistically acceptable...if and only if the property in question is R-related to...basic physical properties (195-96).

But why can a naturalist not believe that there will be many such relations that cannot be specified in advance of science? The philosophical Puritanism that drives this argument comes not from naturalism, but from Stich's own standards of philosophical clarity.

To see how tendentious premise (1) is, consider the obvious authorities. In "Epistemology Naturalized," W. V. Quine[3] observed that in the "new epistemology" he was proposing "we can make free use of empirical psychology" (*ibid.*, p. 24). Similarly, Alvin Goldman[4] explained that although the fathers of epistemology had drawn on psychology, that practice had been rejected by Gottlob Frege and his positivist successors; so in reuniting epistemology with psychology, his new naturalistic approach was a return to the tradition (*ibid.*, pp. 6-7). Given Goldman's account, Stich's connection of naturalism to positivism seems, well, unnatural. Quine and Goldman are concerned with naturalism in epistemology; Stich is considering an ontological thesis. Still, when doing ontology, many naturalists follow Wilfrid Sellars's[5] aphorism: "science is the measure of all things, of what is that it is, and of what is not that it is not" (*ibid.*, p. 173). (Since this position is close to Stich's "open-ended pluralism," the debate threatens to become verbal.) The common thread of these "naturalisms" is the belief that the sciences (not just physics, and especially psychology and biology) matter to philosophy. Although they contain many insightful and important discussions, I think the tendentious conclusions of essays one, three, and five are driven by Stich's own philosophical Puritanism and his odd views of naturalism.

In "Deconstructing the Mind," he aims to show that eliminativists proffered an enthymatic argument, whose missing premise cannot be supplied. The difficulty is that premise one (beliefs, desires, and so forth are the posits of folk psychology) and premise two (folk psychology is false) of the eliminativists' argument can deliver the conclusion (beliefs and desires do not exist) only if buttressed by some account that fixes the reference of terms like 'belief' and 'desire'. He argues that no such account is in the offing. Since one of the likely suspects is a "naturalistic" approach to reference, his portrayal of naturalism needs to be accurate.

---

[3] Reprinted in Hilary Kornblith, ed., *Naturalizing Epistemology* (Cambridge: MIT, 1987).

[4] *Epistemology and Cognition* (Cambridge: Harvard, 1986).

[5] "Empiricism and the Philosophy of Mind," in *Science, Perception, and Reality* (New York: Routledge, 1963).

Stich considers a broad version of what Larry Laudan calls "normative naturalism," an approach that would try to settle the question of eliminativism by appealing to the history of science, in particular, to episodes where theories were so firmly rejected that they took their ontologies down with them and to other cases where ontologies survived the demise of associated theories. Stich does not reduce normative naturalism to physicalism, but he holds this strategy to Puritanical standards. Normative naturalism can succeed only by yielding a demarcation criterion of ontological overthrow versus perseverance: there must be "one normative-naturalist strategy" which would be guaranteed to resolve issues such as eliminativism (66). Again, however, why should naturalists demand a univocal strategy or principle? Presumably, a normative naturalist would examine current psychological theorizing and compare its models and explanatory strategies to earlier psychological theories to try to determine, first, whether any sort of dramatic change has occurred. That judgment would be supported by comparison with other episodes in the history of science. Once these matters had been dealt with—or found to be irresolvable—then the question of eliminativism might be raised. If there is a problem with arguments for eliminativism, it is not the lack of a philosophical theory of reference, but the lack of careful analysis of plausible current *psychological* theories—theories about the capacities that enable human beings to act and think—and their historical predecessors.

Having taken exception to Stich's conclusion that eliminativism is unprovable without an account of reference, let me say that his analysis of the current state of the debate about reference should be required reading. By raising fundamental and long-overdue questions about what a theory of reference is supposed to be a theory of, and what its sources of evidence might be, this essay lays out the parameters for future discussions.

"What is Folk Psychology?" (chapter three) contributes to the attack on eliminativism by arguing that, on some plausible ways of answering this question, folk psychology is not the sort of thing that could be false. Stich and Ravenscroft provide much needed clarification for an expression that has been omnipresent in recent philosophical discussions. They note that "folk psychology" can be understood as a tacit theory used by the folk, or as a theory of the folk by the folk, or as a theory of the folk by the experts, and so on. Again, however, the argument strikes me as flawed in leaving out the naturalist alternative. Seemingly, the interesting claim is not that our grandmothers did not understand the mental, but that the best theories of cognitive psychology have been confused in something like the ways in which J. B. Watson and B. F. Skinner thought nineteenth-century psychological theories were confused. Read as a thesis about the progress of psychology, eliminativism would be a testable, if implausible, claim.

In "Intentionality and Naturalism" (chapter five), Stich and Laurence seek to deflate a philosophical worry closely related to eliminativism, namely, the suspicion that meaning is a "myth." Since intentional notions cannot be "naturalized," we confront an unpalatable "intentional irrealism" (169). Again, the argument proceeds by eliminative induction: there is no reasonable variety of "naturalism" on which (1) it is impossible to capture the intentional naturalistically and (2) it matters that the intentional cannot be so captured. As the authors observe, it cannot matter much that intentional idioms cannot be defined or that intentional things lack essential properties, since these "deficiencies" are widely shared. It also cannot matter that intentional properties do not supervene on very narrow physical bases; but given all nonintentional properties, the intentional must supervene. This would not amount to "naturalizing" the intentional, because the nonintentional is broader than the physical, but so what?

My problem with this argument is that I see no reason why a naturalist should believe that "intentional properties must be *identical with or supervene on* nonintentional properties" (179) or, a fortiori, on physical properties. To settle the question of intentional irrealism, we need to analyze present and future theories of intentional behavior. Since the latter is impossible, discussions of irrealism (or eliminativism) take on an aura of "futurology" (17). What I have been urging is that philosophy of psychology makes real, if less dramatic progress, when it addresses actual psychological theories. In "How Do Minds Understand Minds?" (chapter four), Stich and Nichols patiently and effectively dissect various arguments purporting to show that recent work in psychology constitutes important evidence that the capacity for social understanding is largely a matter of simulation. The essay represents genuine naturalized philosophy of psychology at its best.

PATRICIA KITCHER

University of California/San Diego
Columbia University

INDEX TO VOLUME XCV

INDEX OF SUBJECTS DISCUSSED IN THE JOURNAL

Abstraction

Fine, Kit: Cantorian Abstraction: A Reconstruction and Defense 95 (12): 599-634 '98

Artifacts

Dennett, Daniel C.: Preston on Exaptation: Herons, Apples, and Eggs. 95 (11): 576-80 '98
Preston, Beth: Why Is a Wing Like a Spoon? A Pluralist Theory of Function. 95 (5). 215-54 '98

Belief, Theory of

Taschek, William W.: On Ascribing Beliefs: Content in Context. 95 (7): 323-53 '98

Biology, Philosophy of

Dennett, Daniel C.: Preston on Exaptation: Herons, Apples, and Eggs. 95 (11): 576-80 '98
Matthen, Mohan: Biological Universals and the Nature of Fear. 95 (3): 105-32 '98
Preston, Beth: Why Is a Wing Like a Spoon? A Pluralist Theory of Function. 95 (5): 215-54 '98

Cantor, Georg

Fine, Kit: Cantorian Abstraction: A Reconstruction and Defense. 95 (12): 599-634 '98

Carnap, Rudolf

Katz, Jerrold J.: The Problem in Twentieth-century Philosophy. 95 (11): 547-75 '98

Cognitive Science

Churchland, Paul M.: Conceptual Similarity across Sensory and Neural Diversity. The Fodor/Lepore Challenge Answered. 95 (1): 5-32 '98
Clark, Andy: Time and Mind. 95 (7): 354-76 '98

Color

Katz, Jerrold J.: The Problem in Twentieth-century Philosophy. 95 (11): 547-75 '98

Consensus

Krasnoff, Larry: Consensus, Stability, and Normativity in Rawls's Political Liberalism. 95 (6): 269-92 '98

Contradictions

Priest, Graham: What Is So Bad about Contradictions? 95 (8): 410-26 '98

Conventions

Millikan, Ruth Garrett: Language Conventions Made Simple. 95 (4): 161-80 '98

Cummins, Robert

Preston, Beth: Why Is a Wing Like a Spoon? A Pluralist Theory of Function. 95 (5): 215-54 '98

Dedekind, Richard

Fine, Kit: Cantorian Abstraction: A Reconstruction and Defense. 95 (12): 599-634 '98

Doom

Scott, Alexander D.: Taking the Measure of Doom. By Alexander D. Scott and Michael Scott. 95 (3): 133-41 '98
Scott, Michael: Taking the Measure of Doom By Alexander D. Scott and Michael Scott. 95 (3): 133-41 '98

Emotions

Matthen, Mohan: Biological Universals and the Nature of Fear. 95 (3): 105-32 '98

Ethics

Hurka, Thomas: How Great a Good Is Virtue? 95 (4): 181-203 '98

Expressions

Fiengo, Robert: Names and Expressions. By Robert Fiengo and Robert May. 95 (8): 377-409 '98
May, Robert: Names and Expressions. By Robert Fiengo and Robert May. 95 (8): 377-409 '98

Fodor, Jerry A.

Churchland, Paul M.: Conceptual Similarity across Sensory and Neural Diversity: The Fodor/Lepore Challenge Answered. 95 (1): 5-32 '98

Frege, Gottlob

Fine, Kit: Cantorian Abstraction: A Reconstruction and Defense. 95 (12): 599-634 '98

Functions

Dennett, Daniel C.: Preston on Exaptation. Herons, Apples, and Eggs. 95 (11): 576-80 '98
Preston, Beth: Why Is a Wing Like a Spoon? A Pluralist Theory of Function. 95 (5): 215-54 '98

Gödel, Kurt

Shapiro, Stewart: Proof and Truth: Through Thick and Thin. 95 (10): 493-521 '98

Good

Hurka, Thomas: How Great a Good Is Virtue? 95 (4): 181-203 '98

Habermas, Jürgen

Krasnoff, Larry: Consensus, Stability, and Normativity in Rawls's Political Liberalism. 95 (6): 269-92 '98

Holism

Becker, Kelly: On the Perfectly General Nature of Instability in Meaning Holism. 95 (12): 635-40 '98
Laurence, Stephen: Multiple Meanings and the Stability of Content. By Eric Margolis and Stephen Laurence. 95 (5): 255-63 '98
Margolis, Eric: Multiple Meanings and the Stability of Content. By Eric Margolis and Stephen Laurence. 95 (5). 255-63 '98

Intentionality

McDowell, John: The Woodbridge Lectures 1997: Having the World in View: Sellars, Kant, and Intentionality. 95 (9): 431-91 '98

Intuition

McDowell, John: The Woodbridge Lectures 1997: Having the World in View: Sellars, Kant, and Intentionality. 95 (9): 431-91 '98

Kant, Immanuel

McDowell, John: The Woodbridge Lectures 1997: Having the World in View. Sellars, Kant, and Intentionality. 95 (9). 431-91 '98

Krieger, Martin H.

Scott, Alexander D.. Taking the Measure of Doom. By Alexander D. Scott and Michael Scott. 95 (3): 133-41 '98
Scott, Michael: Taking the Measure of Doom. By Alexander D. Scott and Michael Scott. 95 (3): 133-41 '98

Kripke, Saul

Fiengo, Robert: Names and Expressions. By Robert Fiengo and Robert May. 95 (8): 377-409 '98
May, Robert: Names and Expressions. By Robert Fiengo and Robert May. 95 (8): 377-409 '98
Stroll, Avrum: Proper Names, Names, and Fictive Objects. 95 (10): 522-34 '98

Taschek, William W.: On Ascribing Beliefs: Content in Context. 95 (7): 323-53 '98

Language, Philosophy of

Becker, Kelly: On the Perfectly General Nature of Instability in Meaning Holism. 95 (12): 635-40 '98
Fiengo, Robert: Names and Expressions. By Robert Fiengo and Robert May. 95 (8): 377-409 '98
Katz, Jerrold J.: The Problem in Twentieth-century Philosophy. 95 (11): 547-75 '98
Laurence, Stephen: Multiple Meanings and the Stability of Content. By Eric Margolis and Stephen Laurence. 95 (5): 255-63 '98
Margolis, Eric: Multiple Meanings and the Stability of Content. By Eric Margolis and Stephen Laurence 95 (5): 255-63 '98
May, Robert: Names and Expressions. By Robert Fiengo and Robert May. 95 (8): 377-409 '98
Millikan, Ruth Garrett: Language Conventions Made Simple. 95 (4): 161-80 '98
Stroll, Avrum: Proper Names, Names, and Fictive Objects. 95 (10): 522-34 '98

Lepore, Ernest

Churchland, Paul M.: Conceptual Similarity across Sensory and Neural Diversity: The Fodor/Lepore Challenge Answered. 95 (1): 5-32 '98

Lewis, David

Heller, Mark: Property Counterparts in Ersatz Worlds. 95 (6): 293-316 '98
Millikan, Ruth Garrett: Language Conventions Made Simple. 95 (4): 161-80 '98

Liberalism

Krasnoff, Larry: Consensus, Stability, and Normality in Rawls's Political Liberalism. 95 (6): 269-92 '98

Logic

Priest, Graham: What Is So Bad about Contradictions? 95 (8): 410-26 '98
Shapiro, Stewart: Proof and Truth: Through Thick and Thin. 95 (10): 493-521 '98

Logic, Philosophy of

Fine, Kit: Cantorian Abstraction: A Reconstruction and Defense. 95 (12): 599-634 '98
Katz, Jerrold J.: The Problem in Twentieth-century Philosophy. 95 (11). 547-75 '98
Priest, Graham: What Is So Bad about Contradictions? 95 (8): 410-26 '98
Shapiro, Stewart: Proof and Truth: Through Thick and Thin. 95 (10): 493-521 '98

Lormand, Eric

Becker, Kelly: On the Perfectly General Nature of Instability in Meaning Holism. 95 (12): 635-40 '98
Laurence, Stephen. Multiple Meanings and the Stability of Content. By Eric Margolis and Stephen Laurence. 95 (5): 255-63 '98
Margolis, Eric: Multiple Meanings and the Stability of Content. By Eric Margolis and Stephen Laurence. 95 (5): 255-63 '98

Marcus, Ruth Barcan

Stroll, Avrum: Proper Names, Names, and Fictive Objects. 95 (10): 522-34 '98

Mathematics, Philosophy of

Fine, Kit: Cantorian Abstraction: A Reconstruction and Defense. 95 (12): 599-634 '98

Meaning, Theory of

Becker, Kelly: On the Perfectly General Nature of Instability in Meaning Holism. 95 (12): 635-40 '98
Laurence, Stephen: Multiple Meanings and the Stability of Content. By Eric Margolis and Stephen Laurence. 95 (5): 255-63 '98

Margolis, Eric: Multiple Meanings and the Stability of Content. By Eric Margolis and Stephen Laurence. 95 (5): 255-63 '98

Memo to Authors

95 (1): 51
95 (9): 492

Metaphysics

Heller, Mark: Property Counterparts in Ersatz Worlds. 95 (6): 293-316 '98

Millikan, Ruth Garrett

Dennett, Daniel C.: Preston on Exaptation: Herons, Apples, and Eggs. 95 (11): 576-80 '98
Preston, Beth: Why Is a Wing Like a Spoon? A Pluralist Theory of Function. 95 (5): 215-54 '98

Mind, Philosophy of

Becker, Kelly: On the Perfectly General Nature of Instability in Meaning Holism. 95 (12): 635-40 '98
Clark, Andy: Time and Mind. 95 (7): 354-76 '98
Dennett, Daniel C.: Preston on Exaptation: Herons, Apples, and Eggs. 95 (11): 576-80 '98
Endicott, Ronald P.: Collapse of the New Wave. 95 (2): 53-72 '98
Laurence, Stephen: Multiple Meanings and the Stability of Content. By Eric Margolis and Stephen Laurence. 95 (5): 255-63 '98
Margolis, Eric: Multiple Meanings and the Stability of Content. By Eric Margolis and Stephen Laurence. 95 (5): 255-63 '98
Preston, Beth: Why Is a Wing Like a Spoon? A Pluralist Theory of Function. 95 (5): 215-54 '98

Moore, G. E.

Hurka, Thomas: How Great a Good Is Virtue? 95 (4): 181-203 '98

Names, Proper

Fiengo, Robert: Names and Expressions. By Robert Fiengo and Robert May. 95 (8): 377-409 '98
May, Robert: Names and Expressions. By Robert Fiengo and Robert May. 95 (8): 377-409 '98
Stroll, Avrum: Proper Names, Names, and Fictive Objects. 95 (10): 522-34 '98

Natural Kinds

Khalidi, Muhammad Ali: Natural Kinds and Crosscutting Categories. 95 (1): 33-50 '98

New Books

95 (3): 160
95 (4): 210-14
95 (6): 322

New Books: Translations

95 (11): 598

Notes and News

95 (1): 52

Numbers

Fine, Kit: Cantorian Abstraction: A Reconstruction and Defense. 95 (12): 599-634 '98

Perception

McDowell, John: The Woodbridge Lectures 1997: Having the World in View: Sellars, Kant, and Intentionality. 95 (9): 431-91 '98

Philosophy, Political: *See* Political Theory.

Political Theory

Krasnoff, Larry: Consensus, Stability, and Normativity in Rawls's Political Liberalism. 95 (6): 269-92 '98

Possible Worlds

Heller, Mark: Property Counterparts in Ersatz Worlds 95 (6): 293-316 '98

Preston, Beth

Dennett, Daniel C.: Preston on Exaptation: Herons, Apples, and Eggs. 95 (11): 576-80 '98

Probability

Scott, Alexander D.: Taking the Measure of Doom. By Alexander D. Scott and Michael Scott. 95 (3). 133-41 '98

Scott, Michael: Taking the Measure of Doom. By Alexander D. Scott and Michael Scott. 95 (3): 133-41 '98

Psychology, Philosophy of

Matthen, Mohan: Biological Universals and the Nature of Fear. 95 (3): 105-32 '98

Quantum Mechanics

Endicott, Ronald P.: Collapse of the New Wave. 95 (2): 53-72 '98

Quine, W. V.

Katz, Jerrold J.: The Problem in Twentieth-century Philosophy. 95 (11): 547-75 '98

Rawls, John

Krasnoff, Larry: Consensus, Stability, and Normativity in Rawls's Political Liberalism. 95 (6): 269-92 '98

Representation, Mental

Becker, Kelly: On the Perfectly General Nature of Instability in Meaning Holism. 95 (12): 635-40 '98

Laurence, Stephen: Multiple Meanings and the Stability of Content. By Eric Margolis and Stephen Laurence. 95 (5): 255-63 '98

Margolis, Eric: Multiple Meanings and the Stability of Content. By Eric Margolis and Stephen Laurence. 95 (5). 255-63 '98

Russell, Bertrand

Fine, Kit: Cantorian Abstraction: A Reconstruction and Defense. 95 (12): 599-634 '98

Science, Philosophy of

Endicott, Ronald P.: Collapse of the New Wave. 95 (2): 53-72 '98

Khalidi, Muhammad Ali: Natural Kinds and Crosscutting Categories 95 (1): 33-50 '98

Sellars, Wilfrid

McDowell, John: The Woodbridge Lectures 1997: Having the World in View: Sellars, Kant, and Intentionality. 95 (9). 431-91 '98

Semantics

Becker, Kelly. On the Perfectly General Nature of Instability in Meaning Holism. 95 (12): 635-40 '98

Churchland, Paul M.: Conceptual Similarity across Sensory and Neural Diversity: The Fodor/Lepore Challenge Answered. 95 (1): 5-32 '98

Laurence, Stephen: Multiple Meanings and the Stability of Content. By Eric Margolis and Stephen Laurence. 95 (5): 255-63 '98

Margolis, Eric: Multiple Meanings and the Stability of Content. By Eric Margolis and Stephen Laurence. 95 (5)· 255-63 '98

Speech-Act Theory

Millikan, Ruth Garrett: Language Conventions Made Simple. 95 (4): 161-80 '98

Tarski, Alfred

Shapiro, Stewart: Proof and Truth: Through Thick and Thin. 95 (10): 493-521 '98

Time

Clark, Andy: Time and Mind. 95 (7): 354-76 '98

Truth

Shapiro, Stewart: Proof and Truth: Through Thick and Thin. 95 (10): 493-521 '98

Universals, Biological

Matthen, Mohan: Biological Universals and the Nature of Fear. 95 (3): 105-32 '98

Virtue

Hurka, Thomas: How Great a Good is Virtue? 95 (4) 181-203 '98

Wittgenstein, Ludwig

Katz, Jerrold J : The Problem in Twentieth-century Philosophy. 95 (11): 547-75 '98

INDEX OF AUTHORS

Including Contributing Authors and Reviewers

Becker, Kelly, On the Perfectly General Nature of Instability in Meaning Holism. 95 (12): 635-40 '98

Bromwich, David, Review of Richard Rorty: *Achieving Our Country: Leftist Thought in Twentieth-century America.* 95 (11): 585-90 '98

Churchland, Paul M., Conceptual Similarity across Sensory and Neural Diversity: The Fodor/Lepore Challenge Answered. 95 (1): 5-32 '98

Clark, Andy, Time and Mind. 95 (7): 354-76 '98

Dennett, Daniel C., Preston on Exaptation: Herons, Apples, and Eggs. 95 (11): 576-80 '98

Endicott, Ronald P., Collapse of the New Wave. 95 (2): 53-72 '98

Fiengo, Robert, Names and Expressions. By Robert Fiengo and Robert May. 95 (8): 377-409 '98

Fine, Arthur, Review of Lee Smolin: *The Life of the Cosmos.* By Steven Weinstein and Arthur Fine. 95 (5): 264-68 '98

Fine, Kit, Cantorian Abstraction: A Reconstruction and Defense. 95 (12): 599-634 '98

Hausman, Daniel M., Review of Vivian Walsh: *Rationality, Allocation, and Reproduction.* 95 (8): 427-30 '98

Heller, Mark, Property Counterparts in Ersatz Worlds. 95 (6). 293-316 '98

Herwitz, Daniel, Review of Thierry de Duve: *Kant after Duchamp.* 95 (3): 142-48 '98

Holtman, Sarah Williams, Review of Onora O'Neill: *Towards Justice and Virtue: A Constructive Account of Practical Reasoning.* 95 (6): 317-21 '98

Hurka, Thomas, How Great a Good Is Virtue? 95 (4): 181-203 '98

Katz, Jerrold J., The Problem in Twentieth-century Philosophy. 95 (11): 547-75 '98

Khalidi, Muhammad Ali, Natural Kinds and Crosscutting Categories 95 (1): 33-50 '98

Kitcher, Patricia, Review of Stephen P. Stich: *Deconstructing the Mind.* 95 (12): 641-44 '98

Krasnoff, Larry, Consensus, Stability, and Normativity in Rawls's Political Liberalism. 95 (6): 269-92 '98

Laurence, Stephen, Multiple Meanings and the Stability of Content. By Eric Margolis and Stephen Laurence. 95 (5): 255-63 '98

Margalit, Avishai, Review of Joseph Raz: *Ethics in the Public Domain: Essays in the Morality of Law and Politics.* 95 (2): 97-104 '98

Margolis, Eric, Multiple Meanings and the Stability of Content. By Eric Margolis and Stephen Laurence. 95 (5): 255-63 '98

Matthen, Mohan, Biological Universals and the Nature of Fear. 95 (3): 105-32 '98

May, Robert, Names and Expressions. By Robert Fiengo and Robert May. 95 (8). 377-409 '98

McDowell, John, The Woodbridge Lectures 1997: Having the World in View: Sellars, Kant, and Intentionality. 95 (9): 431-91 '98

Mele, Alfred R., Review of Robert Kane: *The Significance of Free Will.* 95 (11): 581-84 '98

Millikan, Ruth Garrett, Language Conventions Made Simple. 95 (4): 161-80 '98

Morreau, Michael, Review of Isaac Levi: *For the Sake of the Argument: Ramsey Test Conditionals, Inductive Inference, and Nonmonotonic Reasoning.* 95 (10): 540-46 '98

Oliver, Alex, Review of D. M. Armstrong: *A World of States of Affairs* 95 (10): 535-39 '98

Pettit, Philip, Review Essay of Michael Sandel: *Democracy's Discontent. America in Search of a Public Policy.* 95 (2): 73-96 '98

Preston, Beth, Why Is a Wing Like a Spoon? A Pluralist Theory of Function 95 (5). 215-54 '98

Priest, Graham, What Is So Bad about Contradictions? 95 (8): 410-26 '98

Reginster, Bernard, Review of Karl Löwith: *Nietzsche's Philosophy of the Eternal Recurrence of the Same.* 95 (11): 591-97 '98

Rush, Fred L., Jr., Review of Dieter Henrich: *The Unity of Reason: Essays in Kant's Philosophy* and *The Course of Remembrance and Other Essays on Holderlin.* 95 (3): 149-59 '98

Sarkar, Husain, Review of Jarrett Leplin: *A Novel Defense of Scientific Realism* 95 (4). 204-09 '98

Scott, Alexander D., Taking the Measure of Doom. By Alexander D. Scott and Michael Scott. 95 (3): 133-41 '98

Scott, Michael, Taking the Measure of Doom By Alexander D. Scott and Michael Scott. 95 (3): 133-41 '98

Shapiro, Stewart, Proof and Truth: Through Thick and Thin. 95 (10): 493-521 '98

Stroll, Avrum, Proper Names, Names, and Fictive Objects. 95 (10): 522-34 '98

Taschek, William W., On Ascribing Beliefs: Content in Context. 95 (7). 323-53 '98

Weinstein, Steven, Review of Lee Smolin: *The Life of the Cosmos* By Steven Weinstein and Arthur Fine. 95 (5): 264-68 '98

INDEX OF BOOKS REVIEWED

Listed by Author of Item Reviewed

Armstrong, D. M.: *A World of States of Affairs*—Alex Oliver 95 (10): 535-39 '98

de Duve, Thierry: *Kant after Duchamp.*—Daniel Herwitz. 95 (3): 142-48 '98

Henrich, Dieter: *The Unity of Reason. Essays in Kant's Philosophy* and *The Course of Remembrance and Other Essays on Holderlin.*—Fred L. Rush, Jr. 95 (3): 149-59 '98

Kane, Robert: *The Significance of Free Will*—Alfred R. Mele. 95 (11): 581-84 '98

Leplin, Jarrett: *A Novel Defense of Scientific Realism*—Husain Sarkar. 95 (4): 204-09 '98

Levi, Isaac: *For the Sake of the Argument. Ramsey Test Conditionals, Inductive Inference, and Nonmonotonic Reasoning.*—Micheal Morreau 95 (10): 540-46 '98

Löwith, Karl: *Nietzsche's Philosophy of the Eternal Recurrence of the Same*—Bernard Reginster. 95 (11): 591-97 '98

O'Neill, Onora: *Towards Justice and Virtue: A Constructive Account of Practical Reasoning.*—Sarah Williams Holtman. 95 (6): 317-21 '98

Raz, Joseph: *Ethics in the Public Domain: Essays in the Morality of Law and Politics.*—Avishai Margalit. 95 (2): 97-104 '98

Rorty, Richard: *Achieving Our Country: Leftist Thought in Twentieth-century America*—David Bromwich. 95 (11): 585-90 '98

Sandel, Michael: *Democracy's Discontent: America in Search of a Public Policy.*—Philip Pettit. 95 (2): 73-96 '98

Smolin, Lee: *The Life of the Cosmos.*—Steven Weinstein and Arthur Fine. 95 (5): 264-68 '98

Stich, Stephen P.: *Deconstructing the Mind.*—Patricia Kitcher. 95 (12). 641-44 '98

Walsh, Vivian: *Rationality, Allocation, and Reproduction.*—Daniel M [illegible]man. 95 (8): 427-30 '98

UNITED STATES POSTAL SERVICE™

Statement of Ownership, Management, and Circulation

(Required by 39 U S C 3685)

1 Publication Title: THE JOURNAL OF PHILOSOPHY, INC.

2 Publication No: 0 2 8 2 – 0 4 0 X

3 Filing Date: 9/15/98

4 Issue Frequency: MONTHLY

5 No. of Issues Published Annually: 12

6 Annual Subscription Price: $35.00 Individuals
$75.00 Institution

7. Complete Mailing Address of Known Office of Publication (Street, City, County, State, and ZIP+4) (Not Printer)

The Journal of Philosophy, Inc.
1150 Amsterdam Avenue, MC 4972
New York, New York 10027

8 Complete Mailing Address of Headquarters or General Business Office of Publisher (Not Printer)

The Journal of Philosophy, Inc.
1150 Amsterdam Avenue, MC 4972
New York, New York 10027

9 Full Names and Complete Mailing Addresses of Publisher, Editor, and Managing Editor (Do Not Leave Blank)

Publisher (Name and Complete Mailing Address)

The Journal of Philosophy, Inc.
1150 Amsterdam Avenue, MC 4972
New York, New York 10027

Editor (Name and Complete Mailing Address)

Arthur Danto
The Journal of Philosophy, Inc.
1150 Amsterdam Avenue, New York, New York 10027

Managing Editor (Name and Complete Mailing Address)

Michael Kelly
The Journal of Philosophy, Inc.
1150 Amsterdam Avenue, New York, New York 10027

10 Owner (If owned by a corporation, its name and address must be stated and also immediately thereafter the names and addresses of stockholders owning or holding 1 percent or more of the total amount of stock. If not owned by a corporation, the names and addresses of the individual owners must be given. If owned by a partnership or other unincorporated firm, its name and address as well as that of each individual must be given. If the publication is published by a nonprofit organization, its name and address must be stated.) (Do Not Leave Blank.)

Full Name	Complete Mailing Address
	The Journal of Philosophy, Inc.
	1150 Amsterdam Avenue, MC 4972
	New York, New York 10027

11 Known Bondholders, Mortgagees, and Other Security Holders Owning or Holding 1 Percent or More of Total Amount of Bonds, Mortgages, or Other Securities. If none, check here ☐ None

Full Name	Complete Mailing Address
	The Journal of Philosophy, Inc.
	1150 Amsterdam Avenue, MC 4972
	New York, New York 10027

12 For completion by nonprofit organizations authorized to mail at special rates. The purpose, function, and nonprofit status of this organization and the exempt status for federal income tax purposes (Check one)

☒ Has Not Changed During Preceding 12 Months
☐ Has Changed During Preceding 12 Months
(If changed, publisher must submit explanation of change with this statement)

PS Form 3526, October 1994 (See Instructions on Reverse)

13 Publication Name: THE JOURNAL OF PHILOSOPHY

14 Issue Date for Circulation Data Below: SEPTEMBER1998

15 Extent and Nature of Circulation	Average No. Copies Each Issue During Preceding 12 Months	Actual No. Copies of Single Issue Published Nearest to Filing Date
a Total No. Copies (Net Press Run)	4000	3925
b Paid and/or Requested Circulation (1) Sales Through Dealers and Carriers, Street Vendors, and Counter Sales (Not Mailed)	——	——
(2) Paid or Requested Mail Subscriptions (Include Advertisers' Proof Copies/Exchange Copies)	3500	3402
c Total Paid and/or Requested Circulation (Sum of 15b(1) and 15b(2))	3500	3402
d Free Distribution by Mail (Samples, Complimentary, and Other Free)	150	150
e Free Distribution Outside the Mail (Carriers or Other Means)		
f Total Free Distribution (Sum of 15d and 15e)	150	150
g Total Distribution (Sum of 15c and 15f)	3650	3552
h Copies Not Distributed (1) Office Use, Leftovers, Spoiled	350	373
(2) Return from News Agents		
i Total (Sum of 15g, 15h(1), and 15h(2))	4000	3925
Percent Paid and/or Requested Circulation (15c / 15g x 100)	95.89	95.78

16 This Statement of Ownership will be printed in the ______ issue of this publication. ☐ Check box if not required to publish.

17 Signature and Title of Editor, Publisher, Business Manager, or Owner

Managing Editor

Date: 9/15/98

I certify that all information furnished on this form is true and complete. I understand that anyone who furnishes false or misleading information on this form or who omits material or information requested on the form may be subject to criminal sanctions (including fines and imprisonment) and/or civil sanctions (including multiple damages and civil penalties).